Stručni rad

Renato Šoić, Šandor Dembitz
Fakultet elektrotehnike i računarstva, Zagreb
Hrvatska

AUTOMATSKO PREPOZNAVANJE I SINTEZA GOVORA – MOGUĆNOSTI SUSTAVA *SPICE*

SAŽETAK

U radu su opisane mogućnosti sustava SPICE (Speech Processing – Interactive Creation and Evaluation toolkit for new languages) sa stajališta "naivnog" korisnika. Sustav SPICE kreiran je na Sveučilištu Carnegie Mellon i namijenjen je razvoju govornih tehnologija za tzv. necentralne jezike, u koje se ubraja i hrvatski. U radu su opisane osnovne značajke sustava i objašnjeni osnovni principi rada sustava za automatsko prepoznavanje govora i sustava za sintezu govora. Pobliže su opisane faze i pripadni procesi u postupku uhodavanja sustava SPICE. Opisi se temelje na iskustvima u radu sa sustavom, što je rezultiralo izradbom web-sustava za sintezu govora na hrvatskom jeziku.

Ključne riječi: *prepoznavanje govora, sinteza govora, SPICE*

1. UVOD

Računalne govorne tehnologije u načelu bi svakome trebale omogućiti sudjelovanje u prevladavanju jezičnih barijera (Gore, 1998). Nažalost, konstrukcija sustava za obradu prirodnog jezika i govora do sada je zahtijevala prilično velika sredstva. Makar je danas u svijetu registrirano postojanje 6 912 živućih jezika (Gordon, 2005), tradicionalna računalna obrada govora donedavno je bila dostupna samo tzv. centralnim jezicima (Streiter i sur., 2006), koji se dadu nabrojati na prste ruku. Necentralni jezici doskora su bili, zbog ekonomske neisplativosti ulaganja u konvencionalne metode istraživanja i razvoja, prepušteni znatiželji i entuzijazmu malobrojnih stručnjaka, što potvrđuje i najsvježiji hrvatski primjer (Martinčić-Ipšić i sur., 2008). Unatoč svim uspjesima u obradi govora, stvaranje podatkovnih podloga i transfer govornih tehnologija na nove jezike bili su zahtjevan posao, koji su obavljali stručnjaci. Sustav *SPICE* (SPICE, 2009) kreiran je kako bi se prebrodila dosadašnja ograničenja u razvoju govornih tehnologija za necentralne jezike. Ovaj sustav sa svojom javno dostupnom platformom omogućuje i tzv. naivnim korisnicima, tj. korisnicima bez stručna znanja o metodama tvorbe računalnih govornih sustava, da razviju modele za obradu govora u svome jeziku, da skupe podatke potrebne za "hranjenje" tih modela, te da evaluiraju rezultate modeliranja i "hranjenja" s mogućim iterativnim poboljšanjima.

U 2. poglavlju opisani su ciljevi sustava, opće značajke sustava i značajke korisničkog sučelja. U 3. poglavlju opisan je sustav za automatsko prepoznavanje govora. Ukratko su prikazani i Skriveni Markovljevi modeli, dominantne tehnike pri izgradnji sustava za automatsko prepoznavanje govora. Četvrto poglavlje sadrži opis sustava za sintezu govora. Podrobnije su opisani pojedini dijelovi sustava i postupci za sintezu govora. U 5. poglavlju opisan je rad sa sustavom *SPICE*. Tu je također uključen pregled svih bitnih faza i procesa u postupku uhodavanja sustava. Prikaz se temelji na iskustvima u radu sa sustavom *SPICE* (Šoić, 2008), što je u konačnici rezultiralo prvim javnim sustavom za sintezu govora na hrvatskome (Šoić, 2010; *HascheckVoice*, 2010). Šesto poglavlje donosi naš zaključni osvrt na provedena istraživanja i njihove rezultate.

2. ZNAČAJKE SUSTAVA *SPICE*

SPICE (*Speech Processing – Interactive Creation and Evaluation toolkit for new languages*) je *online* sustav namijenjen razvoju govornih tehnologija. Razvijen je na Sveučilištu Carnegie Mellon, a voditelji projekta su Tanja Schultz i Alan W. Black.

Prema funkcijama, sustav se može podijeliti na dva glavna modula: ASR (*Automatic Speech Recognition*), koji omogućuje analizu govora (pretvaranje govora u tekst) i TTS (*Text-To-Speech*), koji omogućuje sintezu govora

(generiranje govora prema pisanom uzorku). U planu je i dodavanje modula koji bi omogućio i prijevod s jednog jezika na drugi na razini teksta – MT (*Machine Translation*) (Schultz, 2006; Schultz i sur., 2007).

Način rada sustava može se opisati pomoću slike 1, na kojoj je prikazan i modul za prevođenje na razini teksta. Govor na nekome jeziku ulazni je podatak u sustav, a ASR modul pretvara ga u tekst. Kad su na raspolaganju tekstualni podaci, MT modul može obaviti prijevod s jednog jezika na drugi jezik. Rezultat su postupka tekstualni podaci na drugom jeziku. Zatim TTS modul na temelju tekstualnih podataka stvara govor. U konačnici se, iz govora na ulazu, dobivaju tekst i govor na izlazu iz sustava (Schultz, 2006).

Slika 1. Sustavi za procesiranje govora
Figure 1. Speech processing system

Opisani postupak promatra se na trima razinama obrade. Prva razina povezuje se s radom na govornim podacima (prepoznavanje i sinteza). Druga razina temelji se na povezivanju govora i teksta – tu se nalaze pravila izgovora. Treća razina povezuje se s obradom tekstualnih podataka, tj. prijevodom. Prednost je opisane podjele povećana prenosivost i iskoristivost sustava. Najbolji primjer jest prva razina obrade: stvaranje kvalitetne tekstualne i govorne baze podataka prilično je zahtjevan posao i poželjno je da se postojeće snimke i fonemska struktura mogu ponovno iskoristiti (Schultz, 2006; Schultz i sur., 2007).

Korisničko sučelje sustava SPICE

Korisničko sučelje sustava *SPICE* prilagođeno je svim potencijalnim korisnicima, od početnika do stručnjaka. Početnicima su na raspolaganju upute

koje ih vode korak po korak kroz izgradnju jezičnih komponenti. Rezultat takva pristupa jest mogućnost skupljanja informacija od vrlo široke skupine ljudi: od uobičajenih korisnika interneta koji nemaju iskustva s alatima za obradu govora, sve do stručnjaka za govor i jezik.

Kako bi sustav bio dostupan širokom krugu korisnika, bilo je bitno osigurati vrlo male hardverske i softverske zahtjeve. *SPICE* ne zahtijeva ništa više od računala sposobnog za "surfanje" internetom. Svi procesi koji se koriste za obradu podataka i izgradnju komponenti sustava odvijaju se na poslužitelju. Korisnik ne zna i ne može vidjeti što se događa u pozadini njegovih radnji. Na taj način postiže se jednostavnost korištenja sustava. U svakom koraku moguće je pregledati log-datoteke kako bi se mogli utvrditi mogući problemi.

Komponenta sustava zadužena za snimanje govora ostvarena je kao *Java applet*. Kad se pokrene, komponenta za snimanje govora korisniku predstavlja redom odabrane rečenice, dobivene analizom tekstualne baze podataka, i snimanje se provodi jednostavnim postupkom. Snimljene datoteke zatim se šalju na poslužitelj, ali se spremaju i lokalno, na korisnikovu računalu, zbog sigurnosti. Pri snimanju govora potrebno je stvoriti profil za svaku osobu koja se snima. Snimanje je moguće obaviti u više navrata (sesija). U slučaju prekida veze s poslužiteljem, aplikacija za snimanje govora prelazi u *offline* način rada. Time je omogućeno pravilno snimanje govora korisnicima s problemima u internetskoj povezanosti (Badaskar, 2008).

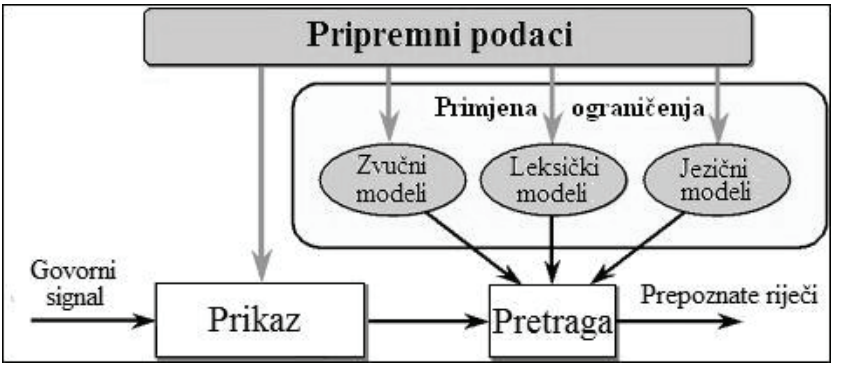
Postupak biranja fonema također je vrlo jednostavan. Korisniku je prikazana IPA (*International Phonetic Alphabet*) tablica koja sadrži sve foneme. Za većinu fonema postoji i zvučni zapis koji omogućuje korisnicima da uspješno definiraju fonemsku strukturu svog jezika, bez poznavanja fonetske abecede. Fonemi se odabiru jednostavnim upisivanjem znakova u kućice pored fonema koji se želi odabrati. I u sljedećoj fazi, određivanju prijelaza iz grafema u foneme, korisniku je ponuđena tablica koja uvelike pojednostavljuje postupak. U ostalim koracima sučelje predstavlja vezu s aplikacijama koje se izvode na poslužitelju, a korisničke akcije svode se na jednostavno pokretanje pojedinih koraka.

3. SUSTAV ZA AUTOMATSKO PREPOZNAVANJE GOVORA (ASR)

Osnovne osobine sustava

Sustavi za automatsko prepoznavanje govora (ASR) vrlo su složeni, i njihov razvoj zahtijeva mnogo resursa. Potrebna su brza računala s mnogo memorije i prostora za pohranu podataka. Osim toga, potrebna je skupina znanstvenika koja uključuje fonetičare, lingviste, računalne stručnjake i matematičare.

Način rada sustava za automatsko prepoznavanje govora može se opisati pomoću slike 2.



Slika 2. Princip rada ASR-a
Figure 2. Working principles of ASR system

Za funkcioniranje sustava za automatsko prepoznavanje govora potrebno je imati što veću bazu podataka teksta i pripadnog snimljenog govora zato što se cijeli sustav uhodava na temelju tih podataka: grade se zvučni i jezični model koji su neophodni za funkcioniranje sustava.

Postupak prepoznavanja govora može se jednostavno opisati kao obrada zvučnog zapisa govora postupkom koji spaja zvučne podatke sa zvučnim i jezičnim modelom, a zatim kao rezultat prepoznavanja generira prepoznatu izjavu.

Digitalni ulazni signal prvo se transformira korištenjem mel-skale. Mel-skala koristi se zato što se na taj način može bolje aproksimirati odziv ljudskog slušnog sustava. Tako se omogućuje bolji prikaz signala i olakšavaju se daljnje operacije, budući da je signal prikazan pomoću niza vektora značajki (engl. *feature vector*) čiji su elementi MFCC koeficijenti (MIT, 2003; Brookes, 2008).

U tijeku faze prepoznavanja, postojeći, pripremljeni zvučni modeli uspoređuju se s obrađenim ulaznim glasom. Viterbijevim se algoritmom (Viterbi, 1967) ostvaruje odabir (dekodiranje), kako bi se promatranom govoru dodijelili najvjerojatniji zvučni modeli. Dekoder prepisuje kontinuirani govor u niz tekstualnih simbola koje aplikacija može izravno obraditi. Cilj je rasporediti simbole u prepoznatljive grupe, usporedbom sa zvučnim modelima. Na kraju postupka slijedi proces najbolje aproksimacije ulaznog govora u formi koju će koristiti nadređena korisnička aplikacija.

Automatsko prepoznavanje govora složeniji je postupak od sinteze govora i zahtijeva mnogo procesorske snage i memorije. Tu nije riječ samo o izgradnji potrebnih modela na temelju skupljenih podataka. Najveći je izazov obavljanje svih potrebnih operacija prilikom testiranja u realnom vremenu.

Sustavima za automatsko prepoznavanje govora problem predstavljaju izrazi koji zvuče (gotovo) isto, ali poruka je bitno različita. Rješenje je u

korištenju baze podataka u kojoj su riječi međusobno povezane s obzirom na kontekst u kojem se zajedno mogu nalaziti (Lieberman i sur., 2005).

Skriveni Markovljev model

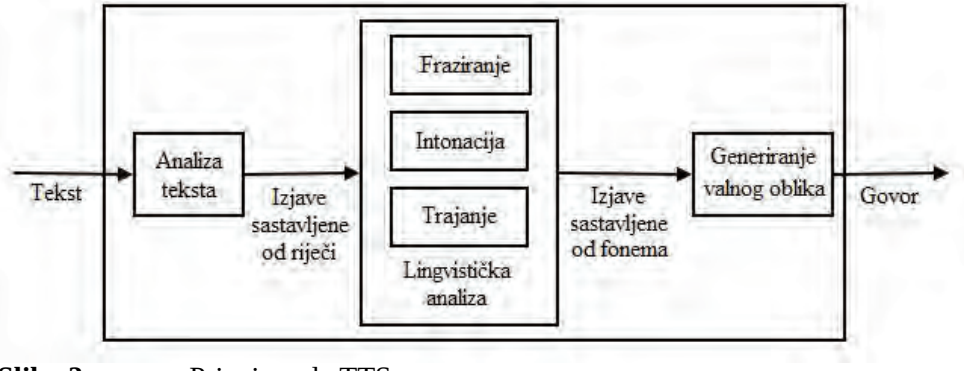
Skriveni Markovljev model (HMM – *Hidden Markov Model*) statistički je model u kojem se modelirani sustav obrađuje kao Markovljev proces s nepoznatim parametrima, a cilj je odrediti skrivene parametre na osnovi vidljivih parametara. Izlučeni parametri modela zatim se koriste u daljnjim analizama. Jedan od razloga zašto se Skriveni Markovljev model koristi u prepoznavanju govora jest činjenica da se govorni signal može prikazati kao stacionarni signal po dijelovima, odnosno može se aproksimirati stacionarnim procesom. Drugi razlog za korištenje Skrivenih Markovljevih modela jest zato što su jednostavni za korištenje i praktični za izračune (Rabiner, 1989).

Skriveni Markovljevi modeli koriste se u kombinaciji s dodatnim tehnikama kako bi se povećala učinkovitost modela. Primjer su normalizacija ulaznog signala i omogućavanje dodatnih stanja u modelu kako bi se ostvarila kontekstna ovisnost fonema. Uz Skrivene Markovljeve modele najčešće se koriste algoritmi koji izračunavaju vjerojatnosti izlaznih nizova na temelju parametara modela, ili izračunavaju vjerojatnosti vrijednosti skrivenih stanja modela. Najčešći su *forward-backward*, Baum-Welch i Viterbijev algoritam (Forney, 1973; Rabiner, 1989; Ephraim i Merhav, 2002).

4. SUSTAV ZA SINTEZU GOVORA (TTS)

Osnovne osobine sustava

Proces pretvorbe teksta u govor (TTS) može se podijeliti na tri osnovna dijela, kao što je prikazano na slici 3. Prvi dio je analiza teksta, zatim slijedi lingvistička analiza i konačno, generiranje valnog oblika.



Slika 3. Princip rada TTS-a
Figure 3. Working principles of TTS system

Analiza teksta

Analiza teksta uključuje postupke čiji je rezultat prepoznavanje riječi u tekstu, tj. nalaženje dobro definiranih metoda izgovora iz leksikona, ili uporaba pravila prijelaza grafema u foneme.

Ulazni tekst prvo se podijeli na manje dijelove, takozvane "izjave", koje se najčešće podudaraju s rečenicama. Podjela teksta u izjave omogućuje sustavu sintetiziranje govora po dijelovima, što znači da će zvučni valovi prve izjave biti raspoloživi prije nego što bi bili kad bi se tekst obrađivao kao cjelina. U većini jezika riječi su odvojene prazninama, a izjave se uglavnom razdvajaju nakon točke, upitnika i uskličnika. U hrvatskom jeziku izjave se odvajaju još i dvotočkom, crticom i točka-zarezom.

Često se u tekstu pojavljuju tokeni koji nemaju izravnu vezu s izgovorom. Tu se pojavljuju dva problema: netrivijalna veza riječi i izraza, te homografi (istopisnice). Primjer su brojevi, koji se drugačije izgovaraju, ovisno o tome koju ulogu imaju. Osim brojeva, posebnu obradu zahtijevaju i simboli za novac, vrijeme, datum, te imena i riječi preuzeti iz drugih sustava pisanja u izvornoj grafiji itd. Normalizacija teksta postupak je koji rješava spomenute probleme, a sastoji se od sljedećih koraka: izoliranje takvih tokena u tekstu, pronalaženje brojeva (ili zapisa poput datuma i vremena) i zamjena riječima u skladu s jezičnim i kontekstualnim pravilima, prevođenje skraćenica u neskraćeni oblik i transkripcija stranih imena, odnosno riječi. Postupak normalizacije brine i o znakovima interpunkcije, bilježi ih i prema potrebi također zamjenjuje riječima (npr. točka unutar *e-mail* adrese).

Poseban problem predstavljaju homografi koji se ispoljavaju kao heterofoni (raznozvučnice), kao npr. u hrvatskome "žena", imenica koju treba izgovoriti "žèna", ako se radi o nominativu jednine, odnosno "žénā", ako se radi o genitivu množine. Homografska raščlamba obično se programski obavlja zajedno s normalizacijom teksta. Značenje sastavnica rečenice procjenjuje se na osnovi poretka i vrsta riječi (parsiranje), što ne jamči uvijek pogađanje konteksta za homografe-heterofone, pa je potrebno koristiti i baze podataka u kojima su pohranjena pravila razlikovanja (Mansoor, 2009).

Lingvistička analiza

Cilj je lingvističke analize pronalaženje izgovora riječi i dodjeljivanje prozodijske strukture izgovoru. Ulazni podaci u lingvističku analizu izjave su sastavljene u fazi analize teksta. Svakoj riječi ulaznih podataka dodjeljuje se fonetska transkripcija, tj. izvršava se prijelaz grafema u foneme. Tekst se zatim dijeli na prozodijske jedinice, poput fraza i rečenica. Prvi je korak u definiranju prozodije određivanje početka i kraja rečenica, čime se određuje visina tona izgovora, ovisno o rečenici i jeziku. Zatim se između rečenica i kod zareza umeću pauze. Ako sustav prepozna fraze, one se naglašavaju slično rečenicama. Izlazni su podaci iz lingvističke analize fonemi, s visinom tona, trajanjem i glasnoćom (Brookes, 2008; Mansoor, 2009).

Stvaranje valnog oblika

Nakon što su određena izgovorna i prozodijska pravila, generira se valni oblik. U ovoj fazi sustav pretvara listu fonema i njihove prozodijske značajke u digitalni zvuk.

SPICE za sintezu govora koristi SPS metodu (*Statistical Parametric Synthesis*). Metoda se pokazala uspješnom u radu s relativno malim bazama podataka snimljena govora, pa je iz tog razloga odabrana za *SPICE*. Sintetizirani govor nije tečan kao kod drugih, razvijenijih metoda, ali je jednostavnije stvoriti jasan glas koji dobro oponaša originalnog govornika (Black i Lenzo, 2007).

5. RAD SA SUSTAVOM

Za postizanje funkcionalnosti ASR i TTS sustava, potrebno je obaviti iste zadatke: stvoriti tekstualnu i govornu bazu podataka, odabrati foneme, definirati pravila prijelaza grafema u foneme, izgraditi jezični i zvučni model, stvoriti izgovorni rječnik i konačno, testirati sustav. Navedeni koraci uhodavanja sustava detaljnije su opisani u nastavku.

Stvaranje tekstualne baze podataka

Kako bi se dobili dobri rezultati, sustavu je potrebno predati što veću količinu teksta. Tekst koji se predaje mora zadovoljavati sljedeća pravila: datoteka u kojoj se tekst nalazi mora biti u *.txt* formatu s UTF-8 načinom zapisa, a svaka rečenica mora biti u vlastitom retku. Jednom kad je skupljena zadovoljavajuća količina teksta, potrebno je pokrenuti analizu teksta kako bi se definirale učestalosti pojavljivanja pojedinih riječi i znakova, prema čemu se stvara skup rečenica prikladnih za snimanje. Na temelju čestotnika teksta, sustav formira skup riječi (leksikon) koje se najčešće pojavljuju. Sljedeći korak predstavlja odabir rečenica prikladnih za snimanje. Sustav u obzir uzima rečenice koje su razumne duljine, sastoje se većinom od riječi iz leksikona i imaju pravilan raspored interpunkcijskih znakova. Jednom kad je stvoren skup prikladnih rečenica, provodi se odabir prema najboljoj fonetskoj zastupljenosti. Završni korak predstavlja formiranje rječnika iz skupa odabranih rečenica. Rječnik se kasnije koristi za izgradnju izgovornog leksikona i govorne baze podataka.

Stvaranje govorne baze podataka

Postupak snimanja vrlo je jednostavan. Nakon što se odabere sesija snimanja i identificira govornik, sve se svodi na snimanje rečenica koje je sustav odredio. Kod postupka snimanja za potrebe ASR sustava, poželjno je snimati što više osoba kako bi se sustav bolje prilagodio. Kod TTS sustava preporuča se snimanje samo jednog govornika, u kontroliranim uvjetima, s jasnim izgovorom.

Odabir fonema

Kako bi sustav inicijalizirao potrebne foneme i mogao ih prepoznavati tijekom analize govornog uzorka, potrebno je definirati fonemsku strukturu jezika. Postupak se temelji na biranju potrebnih fonema iz ugrađenog alata koji sadrži popis svih fonema i pripadne snimljene uzorke. Kad su svi potrebni fonemi odabrani, alat predaje sustavu podatke i stvara se datoteka koja sadrži fonemsku strukturu jezika.

Izgradnja zvučnog modela

Zvučni model (*Acoustic Model*) modelira zvučne jedinice jezika na temelju govornih značajki izlučenih iz zvučnog signala. Izgradnja zvučnog modela odvija se u dva dijela. Prvi dio predstavlja kreiranje *Janus* baze podataka (JANUS, 1998). Tim se postupkom analizira snimljeni govor, i na temelju dobivenih podataka konfigurira se izgradnja ASR-a. Ovisno o količini podataka i broju govornika, odlučuje se koliko puta će se izvršavati drugi dio. Drugi dio odnosi se na pripremu zvučnog modela, a sastoji se od više koraka: inicijalizacija procesa, određivanje oznaka itd.

Izgradnja jezičnog modela

Jezični model predstavlja izračun vjerojatnosti svih mogućih nizova riječi u određenom jeziku. Funkcija jezičnog modela jest pridružiti veće vjerojatnosti pojavljivanja nizovima riječi koji imaju smisla, gramatički su ispravni ili se često pojavljuju, a manje vjerojatnosti besmislenim, gramatički netočnim i rijetkim nizovima riječi. Jednostavnije rečeno, jezični model pokušava predvidjeti sljedeću riječ u govornom uzorku.

Kod postupka izgradnje jezičnog modela, najprije se izračunaju učestalosti pojavljivanja znakova i riječi, a nakon toga gradi se rječnik. U rječnik ulaze sve riječi koje su se u tekstu pojavile više od jednog puta. Slijedi parsiranje izlaznih vrijednosti rječnika i izgradnja modela n-grama.

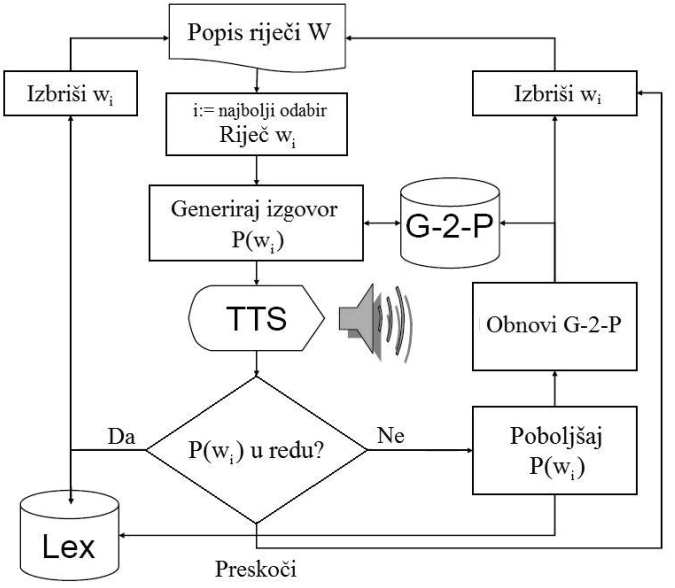
Pravila prijelaza grafema u foneme

Na temelju pravila prijelaza grafema u foneme, sustav "pogađa" točan izgovor riječi u jeziku. Predviđeni izgovor riječi zatim se prikazuje pri stvaranju izgovornog leksikona, i može se mijenjati, ako nije pravilan. Postupak stvaranja pravila prijelaza svodi se na ispunjavanje tablice koja sadrži sve znakove koji su se pojavili u tekstualnoj bazi podataka. Pored svakog znaka potrebno je odabrati vrstu znaka (veliko ili malo slovo, broj, interpunkcijski znak i ostalo) i u polje za unos teksta upisati pripadni fonem.

Stvaranje izgovornog rječnika

Stvaranje izgovornog rječnika postupak je definiranja pravila izgovora – LTS (*Letter To Sound*) pravila. Za izvršavanje te faze potrebno je imati završen jezični i zvučni model. Sustav automatski stvara leksikone na temelju riječi iz tekstualne baze podataka i primjenjujući pravila prijelaza grafema u foneme.

Rezultat postupka su dva izgovorna leksikona i definirana LTS pravila koja će se primjenjivati na nepoznate riječi. Automatski postupak vrlo je jednostavan: sustav odabere riječ iz tekstualne baze podataka, predloži izgovor i omogući slušanje sintetizirane verzije riječi kao diskretnog niza fonema. Korisnik može izgovor prihvatiti ili promijeniti.



Slika 4. Opis stvaranja izgovornog rječnika
Figure 4. Pronunciation dictionary creation

U fazi izgradnje izgovornog rječnika (slika 4), definiraju se i pravila izgovora pojedinih riječi ovisno o kontekstu. *SPICE* zasad ne sadrži alate koji bi to omogućili, a s obzirom da je postupak prilično složen i nije namijenjen širem krugu korisnika, vjerojatno se neće ni naknadno implementirati.

Izgradnja sintetizirana glasa

Operacije koje se izvode pri stvaranju sintetizirana glasa zahtijevaju već korištene podatke. Na temelju snimljena govora, pripadnog teksta i stvorenog izgovornog leksikona, izračunavaju se parametri (*mel cepstrum*) potrebni za izgradnju modela glasa. Iz dobivenih parametara grade se modeli fundamentalne frekvencije i spektralnih koeficijenata u mel-skali – F0 i MCEP CART stablo odluke. Slijedi izgradnja modela trajanja pomoću Skrivena Markovljeva modela, nakon čega preostaje testiranje glasa (Black, 2006; Zen i sur., 2007).

Ocjena uporabljivosti sustava

Tvorci sustava *SPICE* gotovo su u potpunosti ostvarili postavljene ciljeve. Postignuta je željena razina jednostavnosti kako bi se korištenje sustava omogućilo širokom spektru korisnika. Rad je sa sustavom jednostavan i ne zahtijeva posebno informatičko znanje, niti znanje o radu sustava za razvoj govornih tehnologija. Koraci koje treba poduzeti tijekom rada sa sustavom jasni su, i korisnik doslovno treba samo obavljati zadatke kako mu sustav predloži.

S druge strane, postoje i problemi. U slučaju bilo kakva neočekivana ponašanja sustava, korisnik ne može lako pronaći rješenje. Stranice za pomoć objašnjavaju postupke uhodavanja sustava dovoljno dobro, ali kad je riječ o rješavanju problema, sustav nažalost ne nudi nikakvu podršku.

Najozbiljniji propust jest što ne postoji naputak da se ne koriste UTF-8 znakovi pri definiranju fonemske strukture jezika i pravila prijelaza grafema u foneme. Ako korisnik koristi UTF-8 znakove u spomenutim koracima, sustav javlja o pogrešci iz koje se ne može zaključiti u čemu je problem. Osim toga, pri izgradnji zvučnog modela dolazi do pogrešaka u povezivanju seta fonema i izgovornog rječnika. S obzirom na činjenicu da se korisnik ipak nerijetko susreće sa spomenutim poteškoćama, to je nešto što bi se trebalo što prije ispraviti, inače bi rezultati sustava mogli biti ispod očekivanja.

Konačni rezultati rada sa sustavom mogu se ipak pozitivno ocijeniti. Uz skromnu količinu teksta (uzorak od oko 40 000 riječi) moguće je postići funkcionalni sintetizirani glas. Za funkcionalni sustav za automatsko prepoznavanje govora preporučeno je osigurati još veću količinu teksta i snimljena govora.

6. ZAKLJUČAK

S obzirom na resurse potrebne za razvoj sustava za automatsko prepoznavanje govora i sustava za sintezu govora, necentralnim jezicima, u koje nedvojbeno spada i hrvatski, donedavno je bio otežan razvoj takvih sustava. Sustav *SPICE* ponudio je rješenje za prevladavanje dosadašnjih poteškoća jer omogućuje svakom pojedincu koji ima računalo i mogućnost spajanja na internet sudjelovanje u razvoju govornih tehnologija za vlastiti jezik. Ideja tvoraca sustava *SPICE* nije da se razvoj sustavâ u potpunosti prepusti nestručnim pojedincima. Njihova je namjera omogućiti što širem krugu zainteresiranih da sudjeluju u razvoju i postignu određene rezultate u prepoznavanju i/ili sintezi govora za vlastiti jezik. Pri tome se stvaraju odgovarajuće kolekcije podataka u obliku tekstovno-govornih paralelnih korpusa različitih veličina za različite jezike. Vrednovanje uporabljivosti pojedinih segmenata kolekcije, pored onih koji su neki segment kreirali, obavljaju i stručnjaci koji sustav *SPICE* održavaju. Ovakav pristup omogućuje da se s vremenom formiraju velike baze uporabljivih tekstovnih i govornih podataka za neki jezik, koje mogu poslužiti za daljnji stručni razvoj boljih i uspješnijih sustava za prepoznavanje i sintezu govora. Da bi se opisana interakcija stručnjaka okupljenih oko *SPICE-a* i "naivnih" korisnika koji sustav "hrane" podacima o svome jeziku pospješila, poželjna je suradnja lingvističkih, tehničkih i srodnih profila u

oblikovanju reprezentativnih tekstovno-govornih korpusa za *SPICE*. Od te suradnje može profitirati jedino jezik zbog kojega su se okupili.

Iskustva stečena u radu sa *SPICE-om* poslužila su nam da kreiramo *HascheckVoice*, prvi javni sustav za sintezu govora na hrvatskom jeziku s potencijalno neograničenom uporabom vokabulara. *HascheckVoice* je o *SPICE-u* neovisan sustav, no u njegovu razvoju iskorištena su stručna znanja inkorporirana u *SPICE*. Tehnološka podloga im je zajednička (FESTIVAL, 2010; *FestVox*, 2010), s time da nas je autonomno korištenje te podloge riješilo nekih problema s kojima smo se susreli u radu sa *SPICE-om*. Posebna vrijednost autonomnosti mogućnost je kreiranja izgovornog rječnika za hrvatski jezik, sa svim značajkama potrebnim za opis hrvatskog jezika i izgovora, što sâm *SPICE* ne omogućuje u potpunosti.

REFERENCIJE

Badaskar, S. (2008). *Speech interface for SPICE*, radni izvještaj, 6. str., http: //www.is.cs.cmu.edu/11-733/Spice/badaskar.pdf [pristupljeno 29. travnja 2009].

Black, A. W. (2006). CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling. *Proc. Interspeech 2006*, 1762–1765, Pittsburgh, SAD, rujan 2006.

Black, A. W., Lenzo, K. A. (2007). *Building Synthetic Voices*, http://festvox.org/ bsv/, [pristupljeno 29. travnja 2009].

Brookes, M. (2008). *Speech Processing*, http://www.ee.ic.ac.uk/hp/staff/dmb/courses/speech/speech.htm, [pristupljeno 29. travnja 2009].

Ephraim, Y., Merhav, N. (2002). Hidden Markov processes, *IEEE Transactions on Information Theory*, **48** (6), 1518–1569.

FESTIVAL (2010). *The Festival Speech Synthesis System*, http://www.cstr.ed.ac.uk/projects/festival/, [pristupljeno 3. studenog 2009].

FestVox (2010). http://festvox.org/, [pristupljeno 3. studenog 2009].

Forney, G. D. (1973). The Viterbi algorithm. *Proceedings of the IEEE*, **61** (3), 268–278.

Gordon, R. G., Jr. (ur.) (2005). *Ethnologue: Languages of the World*, 15th edition. SIL International, Dallas, SAD. http://www.ethnologue.com/ [pristupljeno 29. travnja 2009].

Gore, A. (1998). Digital declaration of interdependence, in *Remarks from Vice President Al Gore,* 15th International ITU Plenipotentiary Conference, Minneapolis, SAD, www.itu.int/newsarchive/press/PP98/Documents/Statement_Gore.html [pristupljeno 29. travnja 2009].

HascheckVoice (2010). http://hascheck.tel.fer.hr/voice2/voice.html, [pristupljeno 3. studenog 2010].

JANUS (1998). *Janus Speech Recognition Toolkit*,

http://www.cs.cmu.edu/~tanja/Lectures/JRTkDoc/index.html [pristupljeno 29. travnja 2009].

Lieberman, H., Faaborg, A., Daher, W., Espinosa, J. (2005). How to wreck a nice beach you sing calm incense. *Proc. 10th Int. Conference on Intelligent User Interfaces*, 278–280, San Diego, SAD, siječanj 2005.

Mansoor, A. (2009). *How Text-to-Speech Works*, http://project.uet.itgo.com/textto1.htm, [pristupljeno 29. travnja 2009].

Martinčić-Ipšić, S., Ribarić, S., Ipšić, I. (2008). Acoustic modelling for Croatian speech recognition and synthesis. *Informatica*, Litva, **19** (2), 227–254.

MIT (2003). *MIT OpenCourseWare: Automatic Speech Recognition*, http://ocw.mit.edu/OcwWeb/Electrical-Engineering-and-Computer-Science/6-345Automatic-Speech-RecognitionSpring2003/LectureNotes/, [pristupljeno 29. travnja 2009].

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77** (2), 257–286.

SPICE (2009). http://plan.is.cs.cmu.edu/Spice/spice/index.php [pristupljeno 29. travnja 2009].

Streiter, O., Scannell, K. P., Stuflesser, M. (2006). Implementing NLP projects for non-central languages: Instructions for funding bodies, strategies for developers. *Machine Translation*, Kluwer, SAD, **MT-20** (4), 267–289.

Schultz, T. (2006). Rapid language portability of speech processing systems. *ISCA Tutorial and Research Workshop on Multilingual Speech and Language Processing*, Stellenbosch, Južnoafrička Republika, travanj 2006, pozvano predavanje.
http://www.cs.cmu.edu/~tanja/Papers/MULTILING2006-Schultz-Spice.pdf [pristupljeno 29. travnja 2009].

Schultz, T., **Black, A. W., Badaskar, S., Hornyak, M., Kominek, J.** (2007). SPICE: Web-based tools for rapid language adaptation in speech processing systems. *Proc. Interspeech 2007*, 2125–2128, Antwerp, Belgija, kolovoz 2007.

Šoić, R. (2008). *Upoznavanje sa sustavom SPICE*, završni rad, Fakultet elektrotehnike i računarstva, Zagreb, lipanj 2008.

Šoić, R. (2010). *Sinteza hrvatskog govora uporabom sustava Festival*, diplomski rad, Fakultet elektrotehnike i računarstva, Zagreb, lipanj 2010.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, **13** (2), 260–269.

Zen, H., Nose, T., Yamagishi, J., Sako, Sh., Masuko, T., Black, A. W., Tokuda, K. (2007). The HMM-based speech synthesis system (HTS) Version 2.0. *Proc. 6th ISCA Workshop on Speech Synthesis*, 33–46, Bonn, Njemačka, kolovoz 2007.

Renato Šoić, Šandor Dembitz
Faculty of Electrical Engineering and Computing, Zagreb
Croatia

AUTOMATIC SPEECH RECOGNITION AND SYNTHESIS – ABILITIES OF *SPICE* SYSTEM

SUMMARY

This paper describes SPICE system capabilities (Speech Processing – Interactive Creation and Evaluation toolkit for new languages) from a "naive" user viewpoint. The system is created at the Carnegie Mellon University and it is dedicated to the development of speech technologies for the so-called non-central languages. The Croatian language belongs to that group of languages. The paper presents basic characteristics of the SPICE system and explains how the creation of automatic speech recognition (ASR) and text-to-speech (TTS) system for a new language can be performed. SPICE system tuning processes and phases are described in detail. The description was based on the working experience with the system, which resulted in a Web-based Croatian Speech Synthesis System.

Key words: *speech recognition, speech synthesis, SPICE*