Sintetizator govora za slovenščino eBralec

Jerneja Žganec Gros,* Boštjan Vesnicer,* Simon Rozman,† Peter Holozan,† Tomaž Šef‡

* Alpineon razvoj in raziskave, d. o. o.
Ulica Iga Grudna 15, 100 Ljubljana
jerneja.gros@alpineon.si
bostjan.vesnicer@alpineon.si
† Amebis, d. o. o.
Bakovnik 3, 1241 Kamnik
simon.rozman@amebis.si
peter.holozan@amebis.si
‡ Institut Jožef Stefan
Jamova 39, 1000 Ljubljana
tomaz.sef@ijs.si

Povzetek

V članku predstavljamo novi sintetizator govora za slovenski jezik, *eBralec*, ki je prvenstveno namenjen slepim in slabovidnim uporabnikom ter osebam z motnjami branja. Poglavitna prednost *eBralca* v primerjavi s predhodnimi sintetizatorji govora za slovenski jezik je v občutno višji stopnji naravnosti rezultirajočega sintetičnega govora. Ženski glas *eBralec Maja* predstavlja prvi ženski slovenski sintetični glas, ki je bil že dolgo na spisku želja slepih in slabovidnih uporabnikov. V članku predstavljamo zgradbo novega sintetizatorja govora, njegove module ter jezikovne vire, ki so bili uporabljeni pri njegovem razvoju.

The eBralec Speech Synthesis System for Slovenian

A new text-to-speech synthesis system for the Slovenian language, *eBralec*, is presented in the paper. *eBralec* has been developed based on a thorough analysis of user requirements provided by the primary end user group representing blind and visually impaired users. *eBralec* outperforms existing solutions for Slovenian speech synthesis in the output speech quality as it yields close-to-natural sounding output speech. *eBralec* also includes a female voice, which represents the first Slovenian synthetic female voice, which has topped the end user wish lists for a considerable time. In the paper we present the structure of the new speech synthesiser and provide a description of its modules and the underlying language resources.

1 Uvod

V članku predstavljamo novi sintetizator govora za slovenski jezik, *eBralec*. *eBralec* je bil razvit v okviru projekta Knjižnica slepih in slabovidnih in je prvenstveno namenjen slepim in slabovidnim uporabnikom ter osebam z motnjami branja.

Poglavitna prednost *eBralca* v primerjavi s predhodnimi sintetizatorji govora za slovenski jezik, kot so denimo *S-5* (Gros et al., 1997), *Govorec* (Šef, 2002), *Proteus TTS* (Žganec Gros in Žganec, 2008) ter *eSpeak*, je v občutno višji stopnji naravnosti rezultirajočega sintetičnega govora.

Na željo končnih uporabnikov je bila velikost pomnilniškega prostora, potrebnega za namestitev ter delovanje sintetizatorja govora, ohranjena na ravni predhodnih Govorčeve oz. Proteusove. To je narekovalo tudi izvedbeno različico končnega sintetizatorja govora, ki temelji na parametrični predstavitvi zakonitosti govora v slovenskem jeziku. Teh zakonitosti se sintetizator govora nauči samodejno na podlagi obsežnega učnega govornega korpusa, ki je bil posebej posnet v te namene, in ki vključuje relevantne akustične ter prozodijske fenomene, ki so značilni za govorjeno slovenščino.

Nova glasova *eBralca* sta moški glas, *eBralec Renato*, ter ženski glas, *eBralec Maja*. Ženski glas *eBralec Maja* predstavlja prvi ženski sintetični glas, ki je na voljo za slovenščino, in je bil že vrsto let na spisku želja slepih in slabovidnih uporabnikov. Za uporabnike, ki so bili vajeni predhodnikov *eBralca*, so razvijalci v *eBralca* vključili tudi glasove iz *Govorca*, razmišljajo pa tudi o vključitvi Proteusovih glasov.

eBralec bo slepim in slabovidnim uporabnikom znatno olajšal delo z računalnikom, dostop do novic in informacij ter tako omogočil njihovo boljšo e-vključenost v sodobno informacijsko družbo.

V članku predstavljamo zgradbo novega sintetizatorja govora, njegove module ter jezikovne vire, ki so bili uporabljeni pri njegovem razvoju.

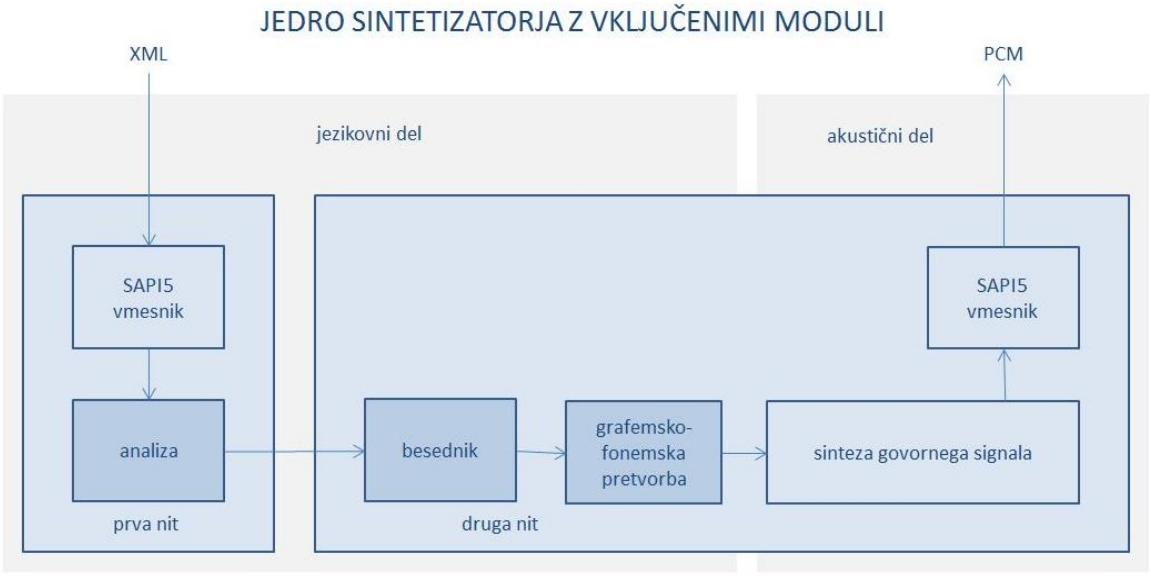
2 Zgradba sintetizatorja govora

Naloga *jedra* sintetizatorja govora oziroma povezovalnega cevovoda je povezovanje sestavnih modulov sintetizatorja govora v enoten proces.

Jedro sintetizatorja govora usklajuje delo posameznih delov sintetizatorja tako, da v ustreznem vrstnem redu vključuje oziroma kliče module sintetizatorja govora. Posamezni moduli oz. faze pretvorbe zaradi pohitritve in večje paralelizacije procesov lahko hkrati delujejo v ločenih nitih (procesorjih ali računalnikih). Zaradi enakomerne porazdelitve procesne obdelave se v prvi niti v trenutni izvedbi izvajajo vsi moduli, potrebni za analizo besedila, v drugi pa ostali moduli, ki so potrebni za nemoteno delovanje sintetizatorje ob izklopu te analize.

Zasnova jedra sintetizatorja govora *eBralec* je prikazana na sliki 1. Moduli, ki jih vključuje jedro *eBralca*, so: jezikovni analizator, besednik, modul za grafemsko-fonemsko pretvorbo in modul za sintezo govornega signala.

Na vhodu in izhodu se jedro sintetizatorja govora lahko poveže z ustreznim vmesnikom, npr. SAPI 5, s pomočjo katerega vhodno besedilo z morebitnimi dodatnimi ukazi spreminja v ustrezen govorni signal.



Slika 1: Shema *jedra* sintetizatorja govora *eBralec*.

Vhodno besedilo sprva obdela *jezikovni analizator*, ki poskrbi za ustrezno predobdelavo vhodnega besedila ter razdvoumljanje izgovornih različic. Rezultat modula za jezikovno analizo je zapis, v katerem so vsebovane vse potrebne informacije o izgovarjavi besed glede na njihovo pozicijo in pomen v vhodnem stavku oziroma povedi.

Modul *besednik* v odvisnosti od vhodnih nastavitev poskrbi za pretvorbo simbolov in števil v besede. Ti elementi so namreč zelo pogost sestavni del besedil, zato je njihovo pravilno izgovarjanje pomembno za razumljivost govora.

Modul »*grafemsko-fonemska pretvorba*« poskrbi za pretvorbo v fonemski zapis (Gros et al., 1997).

Modul za »*sintezo govornega signala*« je zadolžen za oblikovanje prozodije in tvorjenje izhodnega govornega signala.

3 Jezikovna analiza vhodnega besedila

Jezikovna analiza uporablja podatke iz Amebisove *jezikovne baze Ases* (Arhar in Holozan, 2009). Ta za slovenščino v tem trenutku vsebuje več kot 257.000 lem, ki vsebujejo 8,1 milijona oblik, od katerih je 5,7 milijona oblik dodatno opremljenih s podatki o izgovarjavi. Dodatno je za slovenščino v bazi še 36.000 zvez in 8.000 glagolskih predlog. Glagolske predloge podajajo informacije o vezljivosti glagola (Holozan, 2004).

Jezikovni analizator mora narediti razrez besedila na povedi, stavke in besede, potem pa za vsako besedo določiti še ustrezno *lemo* in *oblikoskladenjsko oznako*. *Ases* ločuje leme, ki se različno izgovarjajo, npr. »téma« in »temà« predstavljata dve ločeni lemi.

Že sam razrez vhodnega besedila je lahko težaven. Tako je npr. v primeru »*Videl sem ga. Micka ga je tudi videla*.« treba narediti dve povedi, v primeru »*Videl sem in ga. Micka ga je tudi videla*.« pa le eno. Poleg krajšav so glede tega težavni tudi vrstilni števniki, npr. v primeru »*Bil je na 28. Mednarodnem festivalu*.« Razrez vpliva tako na stavčno intonacijo, kot tudi na branje same kritične besede (*osemindvajset* proti *osemindvajsetem*).

Za branje so predvsem pomembni primeri besed (lahko bi jih imenovali *raznoglasnice*), ki se različno izgovarjajo glede na *pomen v stavku*, in v slovenščini je takih besed zelo veliko.

Take besede so denimo »*je*« (*biti* ali *jesti* ali *osebni zaimek*), »*pol*« (»*Ob pol je pol ljudi šlo na severni pol*.«), »*samo*«, »*celo*«, »*tema*«, »*tako*«, »*mora*«, »*svet*«, »*leti*« (»*Dve leti že leti in leti so vedno daljši*.«), »gori« (»*Gori na gori gori*.«), »hotel« (»Hotel sem hotel.«), tudi »*me*«. Tak primer je npr. tudi »*Vršič*«, kjer je izgovarjava odvisna od tega, ali gre za prelaz ali priimek.

Pri nekaterih glagolih se pri zapisu prekrivata tudi povedni sedanjik in velelnik, ki pa se različno izgovarjata: »*Mati božja prosi za nas*!« proti »*Mati božja, prosi za nas*!«. V slednjem primeru je dodatna težava lahko še to, ali je pisec prav postavil vejico.

Jezikovni analizator, vgrajen v *eBralca*, deluje na podlagi pravil in podatkov iz jezikovne baze *Ases*, pri čemer so osnova glagolske predloge. Analizator je uporabljen tudi v strojnem prevajalniku *Presis* in slovničnem preverjevalniku *Besana*.

Slaba stran zapletenega analizatorja, ki je potreben za razdvoumljanje izgovornih različic besed, je njegova časovna zahtevnost, saj za tekoče branje zahteva razmeroma hiter računalnik. Pokazalo se je, da je analizator v povprečju sicer dovolj hiter, pojavljajo pa se izolirani kritični primeri povedi, v katerih se jezikovna analiza ekstremno podaljša. Najbolj težavne so povedi z velikim številom vejic (npr. dolga naštevanja) in pa daljša zaporedja predložnih zvez, pri katerih se analizator odloča med tem, ali gre za povedkova določila ali za desne prilastke, pri čemer lahko najde zelo veliko število možnih kombinacij.

V *eBralcu* to težavo rešujemo s predčasno prekinitvijo analize, ko se preseže določeno število poskusov, vendar je čas analize v nekaterih primerih še vedno predolg. Slaba stran prekinitve je tudi potencialno nepravilno analizirana poved. Dolgoročna rešitev bo predelava jezikovnega analizatorja, smiselno tudi v smeri izrabe več

jeder procesorja pri analizi, saj trenutno analizator uporablja le eno jedro procesorja, težava pa je, da se v zadnjih letih hitrost računalnikov povečuje bolj z dodajanjem novih jeder kot s samo hitrostjo delovanja jeder (Mattsson, 2014).

4 Tvorjenje govornega signala

Kot smo že omenili v uvodnem poglavju, smo pri razvoju novega sintetizatorja govora upoštevali željo končnih uporabnikov po kompaktni namestitvi. To je narekovalo izvedbeno različico končnega sintetizatorja govora, ki temelji na parametrični predstavitvi zakonitosti govora v slovenskem jeziku. Teh zakonitosti se sintetizator govora nauči samodejno na podlagi obsežne učne govorne zbirke, ki je bila posebej posneta v te namene, in ki vključuje relevantne *akustične* ter *prozodijske* fenomene, ki so značilni za govorjeno slovenščino.

V tem poglavju predstavljamo govorno zbirko, ki smo jo uporabili za učenje parametričnih modelov govora ter postopke modeliranja prozodije in tvorjenja govora s pomočjo prikritih Markovovih modelov.

4.1. Govorna zbirka *eBralca*

Najpomembnejša dejavnika pri snovanju govorne zbirke za potrebe visokokakovostne sinteze govora sta izbira njene vsebine in označevanje posnetkov. Izbira velikosti govorne zbirke je posledica kompromisa med želenim številom variacij glasov oz. njihovim pokritjem na eni strani ter časom in stroški, vezanimi na razvoj na drugi strani. Upoštevati je potrebno tudi čas za kasnejše preiskovanje govorne zbirke in potreben prostor za njeno hranjenje (Amdal in Svendsen, 2005; Hunt in Black, 1996).

Kakovostna sinteza govora zahteva, da ima govorna zbirka pravilno označeno tako identiteto posameznih govornih segmentov kot tudi njihov natančen položaj znotraj zbirke. Običajno samodejnim postopkom za označevanje govorne zbirke sledi »ročno« popravljanje oznak, ki je ne glede na hiter razvoj tehnologije še vedno časovno potratno.

Postopek *zasnove govorne zbirke eBralca* je obsegal naslednje korake (Šef in Romih, 2011):

- ustvari se obsežna tekstovna zbirka besedil, ki pokriva različne zvrsti (dnevni časopis, revije, leposlovje ipd.),
- iz zbirke besedil se odstranijo vse oznake, vezane na oblikovno podobo (glava besedila, tabele ipd.),
- okrajšave, števila ipd. se pretvorijo v polno besedno obliko (normalizacija besedil),
- besedila se pretvorijo v predvideni fonetični prepis (grafemsko-fonemska pretvorba),
- da bi dosegli statistično ustrezno vzorčenje izbranega področja govorjenega jezika, se optimizira obseg zbirke glede na vnaprej pripravljene kriterije z uporabo postopkov požrešnega iskanja,
- izbrane povedi se posnamejo,
- posneto govorno gradivo se fonetično in prozodično označi (samodejno grobo označevanje, fino ročno popravljanje oznak).

4.1.1. Postopek za izbiro povedi

V nadaljevanju bolj podrobno opisujemo postopek za izbiro povedi za snemanje, ki smo mu posvetili veliko pozornosti.

1. *Statistična obdelava besedil*

Statistično obdelamo celoten besedilni korpus in določimo *pogostost pojavljanja posameznih glasov* in *glasovnih nizov* v besedilu. Pri tem dodatno razlikujemo med naglašenimi in nenaglašenimi glasovi ter glasovi, ki se pojavljajo na koncu stavka oz. na mestih zajema zraka ob ločilih (Mihelič et al., 2006). Presledke na drugih mestih lahko prezremo oz. odstranimo, ker je končno besedilo ob branju izgovorjeno povezano.

Vključimo vse stavke (povedne, velelne, vprašalne itd.) in izdelamo statistiko posameznih vrst povedi oz. stavkov.

2. *Izdelava spiska glasovnih nizov* z oceno zaželenosti posameznega niza

V spisek vključimo nabor vseh teoretično možnih kombinacij difonov. Zaradi robustnosti sintetizatorja govora vključimo tudi difone, na katere pri statistični obdelavi nismo naleteli.

V spisek želenih glasovnih nizov vključimo vse trifone, štirifone ter ostale zaželene najpogostejše polifone, na katere smo naleteli pri statistični obdelavi besedil.

Utež oz. ocena zaželenosti niza je odvisna od pogostosti njegovega pojavljanja v besedilu.

3. *Postopek izbire povedi*

Ocenimo doprinos glasovnih nizov za vsako poved iz besedilnega korpusa.

Doprinos povedi je enak vsoti vseh ocen zaželenosti nizov glasov iz spiska želenih glasovnih nizov, ki se v povedi pojavijo.

Doprinos posamezne povedi normiramo z dolžino povedi, izraženo s številom besed v povedi oziroma številom fonemov v povedi.

Določimo takšno utež, da bodo dolžine izbranih stavkov čim bolj ustrezale statistični porazdelitvi dolžin stavkov iz korpusa.

Izberemo poved z najvišjim normiranim doprinosom.

Iz spiska želenih glasovnih nizov odstranimo vse glasovne nize, ki jih izbrana poved vsebuje.

Ponovno ocenimo vsako poved in izberemo najboljšo, glede na novi popravljeni spisek želenih glasovnih nizov.

Postopek ponavljamo, dokler ne dosežemo vnaprej izbranega želenega števila povedi za posamezni sklop obdelave.

4. *Vmesno vrednotenje rezultatov*

Ko obdelamo sklop vnaprej izbranega števila povedi, izdelamo statistiko difonov, trifonov, štirifonov in drugih polifonov, ki jih že pokrivamo: gre za glasovne nize, ki smo jih do takrat že izločili iz zgoraj omenjenega spiska.

5. *Dodatne izboljšave algoritma*

Ker mora zbirka vsebovati vse možne kombinacije difonov, algoritem popravimo tako, da difonom priredimo dodatno težo glede na ostale polifone. Na takšen način bo algoritem na začetku dajal prednost povedim, ki bodo pokrile čim več novih difonov. Predvidoma se vsi difoni pokrijejo že po okoli 100 začetnih dodanih povedih.

Pri trifonih in štirifonih upoštevamo pri robnih glasovih tudi podatek o glasovni skupini, ki ji pripadajo Na primer, štirifon "krak" ne bo doprinesel prav dosti novega v našo zbirko, če ta že vsebuje štirifon "krat", zato oceno koristnosti takega štirifona popravimo navzdol. To lahko naredimo preprosto tako, da v spisek vnesemo

dodatne nize, skupaj z njihovimi frekvencami pojavljanja v korpusu: primer takega štirifona: "k"+"r"+"a"+ "pripornik".

Algoritem za izbiro povedi z različnim uteževanjem izboljšamo tako, da končni nabor vsebuje različne vrste naklonov in raznovrstne povedi: povedne, vprašalne, velelne, enostavne, sestavljene, naštevanje, in podobno. Tako lahko isti korpus učinkovito uporabimo tudi za generiranje prozodijskih parametrov pri sintezi govora.

4.1.2. Snemanje govorne zbirke

Snemanje govorne zbirke je potekalo v studiu RTV Slovenija ob prisotnosti izkušenega tonskega tehnika. Med desetimi profesionalnimi govorci smo izbrali najustreznejši moški in ženski glas. Med branjem besedila so govorci imeli nameščene elektrode laringografa, s katerimi smo spremljali nihanje glasilk za lažje kasnejše označevanje osnovnih period govornega signala.

Samo snemanje je zaradi obsežnosti besedila, ki ga je bilo treba prebrati, trajalo več mesecev. Pri tem so nastavitve opreme ves čas ostale nespremenjene. Pred vsakim snemanjem je govorec poslušal svoje predhodne posnetke, s čimer se je skušalo zagotoviti čim bolj enoten način govora med posameznimi snemalnimi sejami.

Uporabljena govorna zbirka pokriva skoraj vse možne kombinacije difonov in trifonov, ki smo jih identificirali pri analizi dela besedilnega korpusa FidaPLUS[1], ki je obsegal več kot 7 milijonov povedi (približno 30 % povedi v korpusu). Za vsak glas je bilo prebranih 4.000 povedi povprečne dolžine 11 besed.

Sledil je ročni pregled posnetega gradiva, grobo samodejno označevanje mej med glasovi ter časovno zahtevno ročno popravljanje napak.

V primerjavi z obstoječimi govornimi zbirkami, namenjenimi sintezi slovenskega govora (Rojc in Kačič, 2000; Šef, 2002; Mihelič et al., 2006), predstavlja govorna zbirka *eBralec* najbolj obsežno izdelano govorno zbirko za slovensko sintezo govora. V tabeli 1 povzemamo pregledne podatke o novi govorni zbirki.

Velikost besednega korpusa	7.145.345 povedi 77 milijonov besed
Obseg govorne zbirke	*4.000 povedi* 46.785 besed 6 ur 3 min posnetkov za ženski glas 5 ur 33 min posnetkov za moški glas
Število različnih difonov v zbirki	1.883
Število različnih trifonov v zbirki (št. kombinacij v korpusu)	21.369 (24.702)

Tabela 1: Podatki o govorni zbirki *eBralca*. Vsak govorec je posnel 4.000 povedi.

4.2. Modeliranje prozodije in tvorjenje govora z uporabo prikritih Markovovih modelov

Prikriti Markovovi modeli (PMM) so bili vse od prvih poskusov uporabe, ki segajo nekje v sedemdeseta leta prejšnjega stoletja (Jelinek, Baker), pa vse do nedavnega, nepogrešljiva tehnologija na področju samodejne prepoznave govora.

Za razliko od *prepoznave* govora, so se prvi obetavni poskusi *sinteze* govora z uporabo PMM-jev pričeli pojavljati šele v zadnjih letih prejšnjega stoletja. Pionirsko delo na tem področju je opravil Tokuda s sodelavci (Tokuda, 1995). Razlog za razmeroma pozen začetek uporabe lahko pripišemo temu, da se zdijo PMM-ji zaradi svoje statistične narave na prvi pogled neprimerni za nalogo, kot je tvorjenje oziroma sinteza govora. Bistvena razlika med tvorjenjem in prepoznavo govora je namreč ta, da želimo pri prepoznavi iz govora izluščiti le bistvene značilnosti govora, medtem ko želimo pri tvorjenju govora doseči ravno nasprotno, v govoru želimo ohraniti čim več značilnosti, ki so prisotne v naravnem govoru.

Sinteza govora z uporabo PMM-jev ima v primerjavi z bolj klasičnimi postopki tvorbe govora, pri katerih govor tvorimo z »lepljenjem« krajših ali daljših govornih izsekov, nekaj privlačnih prednosti:

- za zadovoljivo kakovost govora potrebujemo razmeroma majhno govorno zbirko (zadošča že ura ali več posnetega govora),
- govorne zbirke ni treba zelo natančno označiti,
- omogoča enovito in sočasno modeliranje akustičnih in prozodičnih lastnosti govora,
- omogoča zgoščen zapis modela govora; za tvorbo govora ni treba hraniti celotne izvorne govorne zbirke,
- omogoča visoko naravnost prozodije tvorjenega govora.

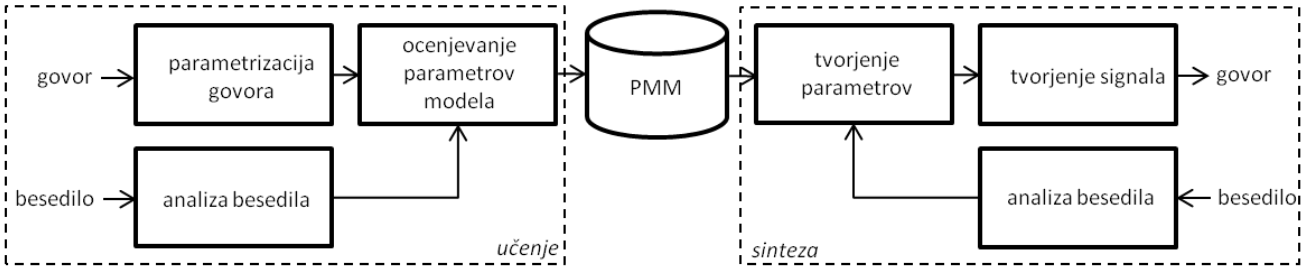
Po drugi strani pa imajo sistemi PMM tudi nekatere slabosti. Govor je lahko na trenutke nekoliko manj razumljiv. Govor ima lahko ponekod značilen »robotski« prizvok, ki je posledica parametrizacije govornega signala.

Od prvih poskusov pa do danes je bilo predlaganih veliko izboljšav in nadgradenj, ki so pripomogle, da je tvorba govora z uporabo PMM-jev postala dobra alternativa bolj uveljavljenim postopkom tvorbe govora, kot je denimo korpusna sinteza govora. Med najpomembnejše tovrstne izboljšave štejemo naslednje:

- postopek generiranja z uporabo dinamičnih parametrov (Tokuda, 10995),
- uporabo kontekstno-odvisnih osnovnih glasovnih enot (Yoshimura et al., 1999),
- združevanje parametrov modela z uporabo fonetičnih pravil (Yoshimura et al., 1999),
- vpeljavo kriterija globalne variance (Toda in Tokuda, 2007),
- uporabo naprednejših metod za parametrizacijo govora, kot je denimo STRAIGHT (Kawahara e tal., 1999),
- uporabo alternativnih statičnih akustičnih modelov (Shannon et al., 2013; Zen in Sak, 2015).

Postopek tvorbe govora z uporabo PMM-jev je sestavljen iz postopka učenja in postopka sinteze. V postopku *učenja* iz večjega števila označenih govornih posnetkov ocenimo parametre modela, v postopku *sinteze* pa naučeni model uporabimo za generiranje govora. Ker surovi zvočni posnetki niso neposredno primerni za gradnjo modela, govor predhodno pretvorimo v zgoščen parametričen zapis. Na sliki 2 je prikazana osnovna shema tvorbe govora z uporabo PMM-jev.

[1] https://sl.wikipedia.org/wiki/FidaPLUS



Slika 2. Shema sistema za tvorjenje govora z uporabo PMM.

V postopku *učenja* želimo izbrati tisto vrednost parametrov statističnega modela, pri kateri bo funkcija verjetja dosegla največjo vrednost. V ta namen po navadi uporabimo učinkovit in teoretično dobro raziskan optimizacijski postopek maksimizacije upanja EM (angl. expectation maximization).

Podoben postopek uporabimo tudi pri *sintezi*, le da v tem primeru na podlagi podanega vhodnega besedila iščemo optimalen niz parametrov govornega signala, medtem ko vrednosti parametrov modela ne spreminjamo. V zadnjem koraku niz parametrov govornega signala pretvorimo v govorni signal.

4.2.1. Tvorjenje govora z uporabo PMM v eBralcu

V nadaljevanju bomo bolj natančno predstavili posamezne korake in različne nastavitve, ki jih uporabljamo za sintezo govora v *eBralcu*. Pri večini korakov smo uporabljali posamezna programska orodja iz sklopa orodij HTS[2], nekatera pomožna orodja pa smo razvili tudi sami.

Govorne zvočne datoteke, katerih frekvenca vzorčenja je znašala 48 kHz, smo najprej pretvorili v nize koeficientov melodičnega kepstra po sledečem postopku. Celoten zvočni signal, ki je ustrezal eni zaključeni stavčni povedi, smo razdelili na 25 ms trajajoče govorne izseke, pri čemer sta se dva sosednja izseka prekrivala v 80-ih odstotkih. Iz vsakega izseka smo izračunali 35 koeficientov melodičnega kepstra, ki smo jim pripeli še logaritem osnovne frekvence, ki smo jo izračunali s postopkom RAPT (Talkin, 1995). Tem 36-razsežnim vektorjem značilk smo dodali še dinamične značilke prvega in drugega reda, tako da smo dobili 108-razsežne vektorje značilk.

Modeli PMM osnovnih govornih enot so imeli pet stanj, topologija pa je bila levo-desna. Trajanje posameznih govornih enot smo modelirali na ekspliciten način – trajanje vsakega stanja modela govorne enote smo opisali z normalno porazdelitvijo. Na podoben način smo modelirali tudi osnovno frekvenco. Posebej smo ocenili tudi globalno varianco značilk na nivoju povedi, za katero se je izkazalo, da lahko pripomore k večji naravnosti govora (Toda in Tokuda, 2007).

Učenje oziroma postopek ocenjevanja parametrov modelov smo izvedli z dobro znanim postopkom Bauma in Welcha, ki predstavlja poseben primer postopka EM.

Pri učenju potrebujemo poleg zvočnih datotek, ki vsebujejo govorne signale (ena datoteka za posamezno stavčno poved), tudi datoteke s fonetičnimi oznakami.

Načeloma je dovolj, če je za vsako poved določen fonetični prepis, še bolje pa je, če so alofoni opremljeni tudi s časovnimi oznakami, ki povedo, kdaj se posamezen alofon v povedi začne in kako dolgo traja. Za dobro kakovost tvorjenega govora je pomembno, da so te oznake čim bolj natančne.

Ker je znano, da so akustične lastnosti posameznega alofona zelo odvisne od okolice, v kateri se alofon nahaja, alofonom pripišemo tudi *kontekst*. Kontekst lahko definiramo poljubno, pomembno pa je, da podaja tiste glasoslovne in jezikoslovne faktorje, ki najbolj vplivajo na akustične lastnosti konkretnega fonema oz. alofona. V našem primeru je kontekst med drugim vseboval naslednje kontekstne faktorje:

- predhodni in sledeči fonem,
- mesto fonema v zlogu,
- mesto zloga v besedi oziroma stavku,
- mesto besede v stavku,
- se fonem nahaja v poudarjenem/nepoudarjenem zlogu,
- razdalja do poudarjenega zloga,
- dolžina prejšnjega/trenutnega/naslednjega stavka,
- stavčni naklon ter
- število zlogov/besed/stavkov v povedi.

Tako definiran kontekst določa osnovno govorno enoto. Idealno bi bilo, če bi za vsako tako govorno enoto lahko naučili lasten PMM, vendar to zaradi kombinatorične eksplozije ni možno. V še tako veliki govorni zbirki bi namreč »videli« le majhen delež vseh takšnih govornih enot.

To težavo rešimo tako, da s pomočjo fonetičnih pravil določimo roje oziroma skupine govornih enot, ki si delijo skupne parametre. Na ta način je mogoče parametre PMM-jev oceniti dovolj robustno. Dodatna prednost deljenja parametrov je tudi ta, da je končni model, ki ga potrebujemo pri sintezi govora, zelo kompakten, četudi je mogoče izvorna govorna zbirka, ki smo jo uporabili za učenje tega modela, zelo obsežna.

5 Zaključek

V prispevku smo predstavili zasnovo in izvedbo novega visokokakovostnega sintetizatorja govora za slovenski jezik. *eBralec* bo slepim in slabovidnim uporabnikom znatno olajšal delo z računalnikom, dostop do novic in informacij ter tako omogočil njihovo boljšo e-vključenost v sodobno informacijsko družbo.

Pri nadaljnjem razvoju *eBralca* imamo še veliko načrtov. Zaradi velike časovne zahtevnosti postopkov jezikovne analize načrtujemo predelavo jezikovnega analizatorja z več stopnjami predelave, od izrazito

[2]HMM-based Speech Synthesis System (HTS), http://hts.sp.nitech.ac.jp.

površinske pa vse do poglobljene jezikovne analize. Razmišljamo v smeri izrabe več jeder procesorja pri jezikovni analizi. Prav tako načrtujemo izboljšave pri sintezi krajših govornih segmentov s kombinacijo korpusne sinteze in PMM.

6 Zahvala

Razvoj *eBralca* je bil delno financiran v okviru projekta Knjižnica slepih in slabovidnih Minke Skaberne. Operacijo je delno financirala Evropska unija iz Evropskega socialnega sklada. Operacija se je izvajala v okviru Operativnega programa razvoja človeških virov, razvojne prioritete "Enake možnosti in spodbujanje socialne vključenosti", prednostne usmeritve "Dvig zaposlenosti ranljivih družbenih skupin na področju kulture in podpora njihovi socialni vključenosti".

7 Literatura

I. Amdal in T. Svendsen. 2005. Unit selection Synthesis Database Development Using Utterance Verification, V: *Zbornik INTERSPEECH 2005*, str. 2553-2556.

Špela Arhar in Peter Holozan. 2009. ASES – leksikalna podatkovna zbirka za razvoj slovenskih jezikovnih tehnologij. V Mikolič (ur.). *Jezikovni korpusi v medkulturni komunikaciji*. Koper: Založba Annales.

Jerneja Gros, Nikola Pavešić in France Mihelič. 1997. Text-to-Speech synthesis: a complete system for the Slovenian language. *CIT*, let. 5, št. 1, str. 11-19.

Peter Holozan. 2004. Uporaba glagolskih predlog pri strojnem prevajanju. V: Zborniku Konference JEZIKOVNE TEHNOLOGIJE 2004, str. 128. Ljubljana.

A. Hunt in A. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. V: *Proceedings of ICASSP 96*, zvezek 1, str. 373-376.

H. Kawahara, I. Masuda-Katsuse in A. de Cheveigné, 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction. *Speech Communication*, zvezek 27, str. 187–207.

P. P. Mattsson. 2014. *Why Haven't CPU Clock Speeds Increased in the Last Few Years?* https://www.comsol.com/blogs/havent-cpu-clock-speeds-increased-last-years/

Aleš Mihelič, Jerneja Žganec Gros, Nikola Pavešić in Mario Žganec. 2006. Efficient subset selection from phonetically transcribed text corpora for concatenation-based embedded text-to-speech synthesis. *Informacije MIDEM*, letn. 36, št. 1, str. 19-24.

Matej Rojc in Zdravko Kačič. 1999. Design of optimal Slovenian speech corpus for use in the concatenative speech synthesis system. V: *Proceedings of the Second international conference on language resources an evaluation*. str. 321-325. Athens. Greece.

M. Shannon, H. Zen in W. Byrne. 2013. Autoregressive models for statistical parametric speech synthesis. *IEEE Trans. Acoust. Speech Lang. Process.*, zvezek 21, št. 3, str. 587–597.

Tomaž Šef. 2002. Sistem GOVOREC za sintezo slovenskega govora. *Elektrotehniški vestnik*, str. 165-170.

Tomaž Šef in Miro Romih. 2011. Zasnova govorne zbirke za sintetizator slovenskega govora Amebis Govorec, V: *Zbornik 14. mednarodne multikonference Informacijska družba*, zvezek A, str. 88-91.

David Talkin. 1995. A robust algorithm for pitch tracking (RAPT). *Speech coding and synthesis*, zvezek 495, str. 518.

T. Toda in K. Tokuda. 2007. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Transactions Inf. Syst.*, zvezek E90-D, št. 5, str. 816–824.

K. Tokuda, T. Kobayashi in S. Imai. 1995. Speech parameter generation from HMM using dynamic features. V: *Proceedings of the ICASSP*, str. 660–663.

T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi in T. Kitamura. 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. V: *Proceedings of the Eurospeech*, str. 2347-2350, september 1999.

H. Zen in H. Sak. 2015. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. V: *Proceedings of the ICASSP*, str. 4470–4474.

Jerneja Žganec Gros in Mario Žganec. 2008. An efficient unit-selection method for concatenative text-to-speech synthesis systems. *CIT*, zvezek. 16, št. 1, str. 69-78.