Umetno tvorjenje slovenskega govora s pomočjo odprto kodnih orodij ter prikritih Markovovih modelov

Tadej Justin[1], France Mihelič[1], Janez Žibert[2]

[1] Univerza v Ljubljani,Fakulteta za elektrotehniko, LUKS, Tržaška 25, 1000 Ljubljana
{tadej.justin, france.mihelic}@fe.uni-lj.si
[2] Univerza na Primorskem, FAMNIT, Glagoljaška 8, 6000 Koper
janez.zibert@upr.si

Povzetek

Na področju tvorjenja umetnega govora (sinteze) se je do danes uveljavilo veliko prosto dostopnih orodji, ki omogočajo gradnjo sistemov za umetno tvorjenje govora (sintetizatorjev). Programska orodja omogočajo s pomočjo vgrajenih uveljavljenih postopkov na uporabniku prijazen način udejanjiti lasten sintetizator govora. Ker so takšni sistemi močno odvisni od jezika, pri tovrstnih orodjih ne najdemo potrebne podpore za gradnjo slovenskega sintetizatorja, kakor tudi ni mogoče zaslediti udejanjenega glasu za slovenski govor. S pomočjo odprto-kodnih programskih orodji za obdelavo govora, obdelavo besedila in njuni analizi, je z nekaj truda veščemu uporabniku omogočeno izdelati tudi jezikovno podporo. S pomočjo analize besedila je mogoče posneti majhno zbirko govora. Z ustrezno urejeno zbirko govora pa je omogočeno, da lahko izgradimo sintetizator govora. V tem prispevku želimo predstaviti prosto dostopna programska okolja za gradnjo sintetizatorjev s pomočjo prikritih Markovovih modelov (PMM), kakor tudi ključne korake za izdelavo slovenske jezikovne podpore ter zajem lastne zbirke slovenskega govora za namen gradnje lastnega sintetizatorja.

Development of slovenian HMM speech synthesis system with the use of open source software

Nowadays in the field of speech synthesis there are many open source frameworks, which allow us to build the text to speech systems. Some of them obtain also the user friendly interface, which allow the inexperienced user to build its own text to speech systems. Since such systems are well dependent on the language, the each language support have to be individually investigated. Currently there is no available language support for Slovenian language and also we did not find any open source Slovenian voice. With some basic knowledge of speech and natural language processing skilled user can build the new language support and based on written language data record small speech database. With obtained database the HMM based text to speech system can easily be build. With this article we want to present the open source frameworks and other helpful software for building the HMM based speech synthesis systems. We present the key steps for building the Slovenian language support, recording small Slovenian speech database and the basic steps, how to obtain the HMM text to speech system with the recorded data.

1. Uvod

Za končne uporabnike je na trgu mogoče najti veliko komercialno dostopnih orodij, ki omogočajo kvalitetno tvorjenje umetnega govora. Poleg takih so prisotna tudi orodja, ki razvijalcem omogočajo enostaven razvoj sintetizatorjev govora. Ker so taki sistemi močno odvisni od jezika, v večini primerov slovenskega jezika ni med podprtimi jeziki. Le malo je razvijalcev, ki se odločijo za gradnjo potrebne jezikovne podpore, ter njeno implementacijo v prosto dostopna programska okolja. Poleg jezikovne podpore je tudi predpogoj za izgradnjo slovenskega sintetizatorja fonemsko in količinsko bogata zbirka slovenskega govora, kar razvijalcem velikokrat predstavlja dolgotrajno ter drago pridobivanje tovrstnih podatkov. To je le nekaj navedb, zakaj med prosto dostopnimi sintetizatorji, še ni takega, ki bi omogočal slovensko sintezo. Avtorji menimo, da ravno prosto dostopna orodja omogočajo poglobljeno poznavanje takih sistemov, saj nudijo uporabniku natančen vpogled tudi v izvorno programsko kodo. V zadnjem času razvojna orodja v veliki meri temeljijo na strokovni javnosti uveljavljenemu postopku za umetno tvorjenje govora s pomočjo prikritih Markovovih modelov (ang. Hidden Markov Models), v nadaljevanju PMM. Njegova uporaba ne zahteva velike procesorske in pomnilniške zmogljivosti računalniških sistemov, kljub temu pa dosega solidne rezultate pri tvorjenju naravnega in razumljivega govora. Med orodji za razvoj tovrstnih sistemov, ki so v zadnjem času med bolj zastopanimi, velika večina temelji na orodju HTS[1] (Yoshimura et al., 1999), ki je izdan kot sistemski popravek za orodje HTK (Young et al., 2006). Orodje ne omogoča za sintezo nujno predpripravo od jezika odvisnih parametrov, vseeno je često uporabljen kot glavna komponenta za izgradnjo sintetizatorjev govora s pomočjo PMM. Med prosto dostopnimi orodji, ki imajo vgrajeno orodje HTS, sta nedvomno najbolj zastopana Festival[2] in Mary TTS[3] (Schröder et al., 2011). Prvega so razvili raziskovalci centra za raziskovanje govornih tehnologij na univerzi v Edinburgu, drugi pa je delo skupine raziskovalcev v laboratoriju za jezikovne tehnologije na nemškem raziskovalnem inštitutu za umetno inteligenco DFKI. S poznavanjem komponent tako enega kot drugega je mogoče udejanjiti tudi jezikovno podporo, ki je predpogoj za gradnjo sintetizatorja s pomočjo PMM.

V tem prispevku želimo spodbuditi razvoj odprto kodnih slovenskih sintetizatorjev govora, kakor tudi predstaviti potrebne korake za gradnjo jezikovne podpore za slovenski jezik. S pomočjo analize besedila nad slovenskim leposlovjem, predstavljamo tudi enostavno snemanje fonetično bogatih povedi, s katerimi lahko slovenski go-

[1] http://hts.sp.nitech.ac.jp/
[2] http://www.festvox.org/festival
[3] http://mary.dfki.de

vorci enostavno posnamejo govorno podatkovno zbirko ter s pomočjo prej naštetih orodji zgradijo lastni sintetizator govora.

2. Gradnja slovenske jezikovne podpore za sintezo govora

Ključna za izgradnjo sintetizatorja je ustrezno pridobljena in označena govorna zbirka. V tem prispevku se osredotočamo na pridobivanje ter analizo besedila lektorirane slovenske pisane besede.

2.1. Pridobivanje slovenskega besedila

V splošnem nam je pisana beseda sama po sebi na voljo na svetovnem spletu. Z razvojem spletnih programskih jezikov so se uveljavili tudi spletni "čitalci" (ang. parsers), ki predstavljajo nepogrešljiv pripomoček pri zbiranju besedila iz svetovnega spleta. Vseeno se je potrebno vprašati o kakovosti pridobljenega slovenskega besedila ter nenazadnje tudi o intelektualni lastnini, avtorstvu in z njim povezanim pravom.

Pri gradnji obsežnejših zbirk besedila iz svetovnega spleta velikokrat pridobimo nekakovostne podatke, ali pa take, ki jih je potrebno ročno obdelati. Da bi si prihranili dragocen čas namenjen ročnemu pregledovanju zbirke besedil, smo se odločili, da besedilo za naše potrebe pridobimo s pomočjo slovenskega leposlovja. Tako besedilo predstavlja lektorirano besedilo, kjer ob enostavnih pravilih izločanja tujih besed, znakov ter številk, lahko pridobimo dovolj dober čistopis, ki nam nudi osnovo za analizo slovenske pisane besede,hkrati pa predstavlja tudi zbirko za pridobivanje fonetično bogatih sklopov povedi, s katerimi lahko posnamemo dovolj bogato fonetično zbirko slovenskega govora za namen gradnje sintetizatorja.

2.2. Analiza pridobljenih povedi

Pridobljen čistopis smo razdelili v tri skupine. Prva zajema povedne povedi (zaključijo se s piko) ter nedokončane povedi (zaključijo se s tremi pikami). Druga vzklične povedi (zaključijo se s klicajem) ter tretja vprašalne povedi (zaključijo se z vprašajem). Taka razdelitev je bila vnaprej predvidena, saj si pri gradnji sintetizatorja želimo zajeti tudi raznoliko stavčno intonacijo, ki omogoča, da je umetni govor bolj naraven. Zato moramo pri zajemu zbirke govora vključiti tudi primerno število povedi vseh treh skupin. Da bi iz čistopisa avtomatsko pridobili čimbolj kakovostne povedi za snemanje zbirke govora, smo izločiti tudi povedi, ki so prekratke ali predolge. Za prve velja, da niso fonetično dovolj bogate. Pri drugih pa je mogoče našteti dva vzroka za njihovo izločitev. Prvi je pogojen z natančnostjo avtomatske poravnave (ang. forced-aligment) govornih enot v akustičnem signalu s fonemskim prepisom besedila, saj je ob predolgih povedih podvržen večji napaki. Drugi vzrok pa izhaja iz uporabniku prijaznega zajema zbirke govora, saj neprofesionalno branje enostavnejših povedi pripomore k boljšemu razumevanju ter posledično boljšemu izražanju. Celoten skupek besedila smo obdelali, tako da smo izločili povedi, ki obsegajo manj kot tri in več kot petnajst besed. Količina povedi v posameznih skupinah je prikazana v tabeli 1.

2.3. Fonemska analiza povedi

Vez med pisanim besedilom (grafemi) ter akustičnim signalom lahko predstavimo z zapisom izgovorjene besede s pomočjo fonetike. Plod dolgoletnega dela sodelavcev v Laboratoriju za umetno zaznavanje, sisteme in kibernetiko na Fakulteti za elektrotehniko, Univerze v Ljubljani, je tudi program za avtomatsko fonemsko grafemsko pretvorbo, ki jo je mogoče uporabljati za namen sinteze slovenskega govora, kakor tudi za potrebe sorodnega področja razpoznavanja slovenskega govora (Gros, 1997). Sestoji iz fonetičnih pravil in skrbno ročno pridobljenega slovarja z več kot 35.000 besedami. S pomočjo programa tako pridobimo fonemski zapis posameznih besed. Za potrebe tega prispevka smo realizirali pretvorbo s 43-timi fonemi ter jim dodali oznako za premor.

V kolikor ne bi razpolagali z avtomatsko grafemsko fonemsko pretvorbo, bi jo lahko z uporabo naprednih algoritmov za obdelavo jezika, ki jih omogočata tudi programska okolja Mary TTS in Festival pridobili avtomatično. Predpogoj za učenje odločitvenih dreves za avtomatsko odločanje nad pravili fonetike je skrbno pripravljen grafemsko fonemski slovar z nekaj več tisoč vnosi ter skrbno določene značilke posameznih fonemov jezika.

Pri gradnji sintetizatorjev s pomočjo PMM, se za modeliranje običajno uporablja trifonske govorne enote. Ti predstavljajo skupek treh povezanih monofonov. S tem pa se število vseh možnih govornih enot bistveno poveča in v našem primeru znaša 85184 vseh možnih trifonov. V kolikor si želimo, da udejanjimo kvaliteten sintetizator, je pomembno, da razpolagamo z zbirko govora, ki ima čim večjo raznolikost možnih trifonov. V tabeli 1 lahko preverimo tudi trifonsko zastopanost podskupin čistopisa slovenskega leposlovja. S pomočjo trifonske zastopanosti posameznih sklopov povedi lahko pridobimo lestvico najbolj pogostih trifonov v obravnavani zbirki slovenskega besedila, kar lahko s pridom uporabimo pri izbiri manjšega števila fonemsko bogatih povedi za snemanje zbirke slovenskega govora.

2.4. Izbira fonemsko bogatih povedi za snemanje govorne zbirke

Izbira manjšega nabora povedi v namen snemanja govornih signalov, ki so najbolj značilne za obravnavan jezik in so hkrati tudi fonemsko bogate ni enostavna naloga. V literaturi se pojavljata dva pristopa. Prvi temelji na naključni izbiri povedi, toliko časa, da skupek povedi zajame vse osnovne govorne enote. Drugi se obračajo na difonsko analizo ter difonsko prozodično porazdeljenost. Pri zadnjem mora uporabnik sam definirati končno število povedi, ki jih želi pridobiti, kakor tudi mora vsakemu monofonu pripisati fonemske značilnosti.

V literaturi (Yamagishi et al., 2010) je možno zaslediti, da je mogoče udejanjiti sintetizator s pomočjo PMM z govorno zbirko s skupno dolžino manj kot 15 minut govora istega govorca. Pri čemer mora govorna zbirka vsebovati vse monofone, obenem pa je potrebno težiti tudi k čim večji trifonski zastopanosti. Da bi navedeno preverili tudi za slovenski jezik, smo povedi, s katerimi smo posneli govorno zbirko izbrali na naslednji način. Najprej smo vsak sklop (povedne, vzklične in vprašalne) povedi dodatno ure-

Tabela 1: Analiza čistopisa zbirke povedi slovenskega leposlovja

Analizirana postavka	Povedne povedi	Vprašalne povedi	Vzklične povedi	Skupaj
Št. povedi	166223	14552	9891	190666
Št. edinstvenih besed	106507	12462	17811	112126
Št. vseh besed	1434211	66694	98548	1599453
Št. edinstvenih trifonov	28218	15336	17591	28673
Procent zastopanosti trifonov	33,12	18,00	20,65	33,66
Št. vseh trifonov	6580055	280368	432537	7292960

Tabela 2: Analiza sklopa izbranih 200 povedi za snemanje zbirke slovenskega govora

Analizirana postavka	Povedne povedi	Vprašalne povedi	Vzklične povedi	Skupaj
Št. povedi	160	20	20	200
Št. edinstvenih besed	115	115	111	102
Št. vseh besed	1476	152	139	1767
Št. edinstvenih trifonov	3568	610	577	4004
Procent zastopanosti trifonov	4,19	0,72	0,68	4,7
Št. vseh trifonov	7014	707	605	8376

dili v pod-sklope glede na količino besed. Vsaki povedi smo pripisali monofonsko zastopanost in jih uredili po velikosti od največje do najmanjše zastopanosti. Iz vsakega pod sklopa smo proporcionalno glede na celotno zastopanost posamezne skupine povedi v zbirki naključno izbrali 160 povednih povedi, 20 vzkličnih povedi ter 20 vprašalnih povedi. Pri tem smo upoštevali, da smo iz vsakega pod sklopa naključno izbirali le v prvi polovici (bolj zastopane) monofonske zastopanosti. Zastopanost sklopa povedi prikazuje tabela 2.

3. Zajem zbirke govora

Za snemanje govora na podlagi besedilne zbirke smo uporabili program Red Start, ki je del programskega orodja Mary TTS. Program omogoča snemanje predvidenih povedi tako, da uporabniku nudi velik izpis povedi, omogoča enostaven izris posnetega akustičnega signala v časovnem in frekvenčnem prostoru, detekcijo šuma, ter večkratno ponavljanje snemanja izbrane povedi. S snemanjem zbirke smo pridobili približno 10 min govora.

4. Splošen sistem PMM sistema za umetno tvorjenje govora

Za udejanjanje lastnega sintezizatorja s pomočjo tehnike PMM je potrebno poznavanje orodja HTS ter njegovih komponent. V splošnem sestoji iz treh ločenih komponent, ki si sledijo po vrstnem redu; analiza govornega signala, učni proces sintetizatorja ter generiranje akustičnega signala (sinteza).

4.1. Priprava govornih signalov za gradnjo PMM sintetizatorja

Vsak akustični signal, ki je namenjen gradnji PMM sintetizatorja mora biti opremljen z grafemskim prepisom. S pomočjo grafemsko fonemske pretvorbe lahko pridobimo tudi fonemski zapis akustičnega signal. Za uspešno gradnjo PMM sintetizatorja moramo priskrbeti tudi časovno poravnavo med fonemi in akustičnim signalom. V kolikor imamo na razpolago ročno označeno fonemsko poravnano, lahko pričakujemo boljšo kvaliteto sintetiziranega govora, kar nakazuje, da natančna poravnava fonemov in akustičnega signala igra pomembno vlogo pri gradnji PMM sintetizatorja. Velikokrat ne razpolagamo z ročno označeno zbirko govornih signalov. V takem primeru moramo uporabiti avtomatsko označevanje trajanja posameznih fonemov vsakega govornega signala v zbirki. Postopki za avtomatično označevanje niso tako natančni kot ročna poravnava, vendar je natančnost pri primerno obsežnem govornem signalu zadovoljiva. Avtomatično poravnavo govorne zbirke lahko izvedemo z orodjem EHMM (Prahallad et al., 2006) ali kar neposredno z uporabo orodja HTK.

Ko pridobimo fonemsko poravnavo z govornim signalom, lahko pričnemo s potrebnim luščenjem značilk govornega signala. Moderni sistemi omogočajo luščenje spektralnih značilk, kot tudi značilk govornega trakta. Najbolj pogosto uporabljene spektralne značilke so koeficienti melodičnega kepstra (ang. mel frequency cepstral coefficinets, MFCC). Pridobimo jih lahko s pomočjo orodja SPTK (Takuda et al., 2009) ali ESP (Black et al., 2003). Za pravilno modeliranje osnovne frekvence potrebujemo tudi ustrezna orodja. Za to je na voljo več prosto dostopnih programov, v našem delu smo uporabili orodje Praat (Boersma and Weenink, 2001).

4.2. Gradnja PMM sintetizatorja

Učni proces je podoben učnemu procesu pri razpoznavanju govora. Spektralne značilke ter značilke vzbujanja (kot na primer logaritem osnovne frekvence ter njene dinamične zančilke ali trajanje posamezne osnovne enote) se pridobivajo iz zbirke govornih posnetkov in so modelirane z množico kontekstno odvisnih PMM. Parametre PMM modelov nato na podlagi teh značilk določimo po kriteriju maksimalnega verjetja (ang. maximum likelihood), v nadalje-

vanju ML,

$$\hat{\boldsymbol{\lambda}} = arg\,max_{\lambda}\,\{p(\boldsymbol{O}|\boldsymbol{W},\lambda)\}, \quad (1)$$

kjer je $\boldsymbol{\lambda}$ vektor parametrov modela,$\boldsymbol{O}$ vektor učnih podatkov in $\boldsymbol{W}$ vektor zaporednih manjših besednih enot, ki pripadajo vektorju O. Parametri govora, $o = o_1, ..., o_T$, se tvorijo s pomočjo vektorja ocenjenih modelov $\hat{\boldsymbol{\lambda}}$ za dan vektor manjših besednih enot, ki bodo sintetizirane, ω, na način, da bodo izhodne verjetnosti največje

$$\hat{\boldsymbol{o}} = arg\,max_{o}\,\{p(\boldsymbol{o}|\omega,\hat{\boldsymbol{\lambda}})\}. \quad (2)$$

Modeliranje trajanja govornega segmenta je enostavnejše, saj v splošnem lahko na podlagi informacije koliko časa ostaja PMM v nekem stanju ocenimo trajanje. Trajanje lahko enostavno ocenimo iz matrike porazdelitve verjetnosti prehodov med stanji. Proces učenja smo v tem prispevku izvedli s pomočjo orodja HTS (Yoshimura et al., 1999).

4.3. Izvedba PMM sintetizatorja s tehnikami prilagajanja

Z razvojem tehnik prilagajanja (ang. adaptation) govorca na različna akustična okolja pri razpoznavanju govora, so se te tehnike začele uporabljati tudi pri sistemih za umetno tvorjenje govora. Tako moderni sistemi za umetno tvorjenje govora s pomočjo PMM omogočajo najprej gradnjo splošnega modela posameznega jezika, ki zajema posplošen akustični model, posplošen model trajanja ter posplošen model vzbujanja. S tehnikami kot sta postopek maksimizacije posteriorne porazdelitve (ang. maximum posteriori, MAP) ter postopek linearne regresije z maksimalnim verjetjem (ang. constrained, maksimum likelihood regression, CMLLR) je mogoče vplivati na parametre PMM splošnega modela govora z namenom prilagajanja parametrov splošnih modelov na posameznega govorca. S takim načinom pridobimo umetno tvorjen govor, ki je podoben lastnemu govoru govorca, oziroma posnetkov, ki jih imamo na razpolago za prilagajanje, pri čemer izrabimo tudi vse akustične značilnosti splošnega modela govora. Že z relativno majhno količino posnetkov namenjenih za prilagajanje je mogoče prilagoditi splošni modela jezika. Na tem mestu moramo opozoriti, da v kolikor fonemi v govoru niso zastopani, se tudi slabše prilagodijo na splošni model, kar se odraža tudi pri sintezi. S pomočjo demonstracijskih programov na spletni strani skupine HTS, smo tako izvedli tudi sintetizator na način, da smo adaptirali posneto govorno zbirko na splošen model, ki smo ga izvedli s pomočjo petih govorcev (1 ženska govorka, 4 moški govorci) zbirke VNTV (Žibert and Mihelič, 2000).

4.4. Sinteza govora

Proces sinteze je obrnjen proces razpoznave govora. Vhod v postopek predstavlja niz osnovnih govornih enot. Na podlagi tega niza se ustrezni modeli povežejo v verigo. Nato se izbere najbolj verjetna pot (niz stanj) skozi to verigo glede na porazdelitve verjetnosti trajanj posameznih stanj. Ta najbolj verjetna veriga odda niz vektorjev v katerih so združene značilke vzbujanja ter govornega trakta. Iz poteka značilk vzbujanja se nazadnje tvori še vzbujanje, ki ga vodimo na vhod filtra, ki smo ga določili s pomočjo spektralnih značilk. Na izhodu filtra dobimo sintetiziran govor (Yoshimura et al., 1999).

5. Zaključek

Čeprav v prispevku nismo ovrednotili udejanjenih sistemov za umetno tvorjenje govora, smo pokazali, da je mogoče s pomočjo PMM sinteze z relativno majhno količino podatkov in pravilno izbiro slovenske pisane besede za gradnjo govorne zbirke izdelati sintetizator slovenskega govora. V nadaljnjem delu bomo nedvomno ovrednotiti udejanjene sisteme ter preizkusili tudi naprednejše algoritme za pridobivanje fonemsko bogatih sklopov povedi v namen snemanja ustreznih govornih zbirk, primernih za sintezo govora. Obenem pa bomo poizkusili določiti še priporočljive najmanjše količine govornega materiala, s katerim lahko še udejanjimo slovenski sintetizator govora s pomočjo PMM.

6. Literatura

A. W. Black, Taylor P., Caley R., Clark R., in S. King. 2003. The edingburgh speech tools library. Tehnično poročilo.

Paul Boersma in David Weenink. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Jerneja Gros. 1997. *Samodejno tvorjenje govora iz besedil*. Doktorsko delo, Fakulteta za elektrotehniko, Univerza v Ljubljani.

Kishore Prahallad, Alan W Black, in Ravishankhar Mosur. 2006. Sub-phonetic modeling for capturing pronunciation variation in conversational speech synthesis. V: *in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*.

Marc Schröder, Marcela Charfuelan, Sathish Pammi, in Ingmar Steiner. 2011. Open source voice creation toolkit for the mary tts platform. V: *Proceedings of Interspeech 2011*. ISCA.

K. Takuda, Masuko T., Koishida K., Sako S., Zen H. Imai S., in Kobayashi T. 2009. Speech signal processing toolkit (sptk). Tehnično poročilo.

Junichi Yamagishi, Bela Usabaev, Simon King, Oliver Watts, John Dines, Jilei Tian, Yong Guan, Rile Hu, Keiichiro Oura, Yi-Jian Wu, Keiichi Tokuda, Reima Karhila, in Mikko Kurimo. 2010. Thousands of voices for hmm-based speech synthesis: analysis and application of tts systems built on various asr corpora. *Trans. Audio, Speech and Lang. Proc.*, 18(5):984–1004, July.

Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, in Tadashi Kitamura. 1999. Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. V: *EUROSPEECH'99*, str. –1–1.

S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, in P. C. Woodland. 2006. *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK.

Janez Žibert in France Mihelič. 2000. Slovenian weather forecast speech database. V: *Proc, SoftCOM*, str. 199–206.