Vrednotenje na prikritih Markovovih modelih temelječega sistema za umetno tvorjenje slovenskega govora

Boštjan Vesnicer, France Mihelič, Nikola Pavešić

Laboratory of Artificial Perception, Systems and Cybernetics
Faculty of Electrical Engineering
University of Ljubljana
Tržaška 25, SI-1000 Ljubljana
{bostjan.vesnicer,france.mihelic,nikola.pavesic}@fe.uni-lj.si

Povzetek

Predstavljamo nov sistem za sintezo slovenskega govora, ki temelji na teoriji prikritih Markovovih modelov. Rezultati izvedenih subjektivnih in objektivnih testov kažejo, da je kakovost sintetiziranega govora v smislu naravnosti v primerjavi s predhodno razvitim difonskim sintetizatorjem na višjem nivoju.

Evaluation of the Slovenian HMM-Based TTS System

A new HMM-based speech synthesis system for Slovenian language is presented. The quality of synthesized speech has been assessed by subjective and objective tests. The results show that the new system outperforms our previously developed diphone-based waveform concatenation synthesizer in terms of naturalness and general impression.

1. Uvod

Razvoj na področju umetnega tvorjenja govora je prišel do točke, ko sintetizatorji govora dosegajo zelo visoko stopnjo razumljivosti, problem pa je v naravnosti umetnega govora, ki za številne namene še vedno ni zadovoljiva. V zadnjem času je bilo predvsem zaradi želje po večji naravnosti moč opaziti znaten premik od difonske sinteze h korpusni sintezi (Campbell in Black, 1996). Prednost korpusne sinteze pred klasično difonsko sintezo je v tem, da imamo pri prvi v nasprotju s slednjo na razpolago več istovrstnih govornih enot (npr. difonov), med katerimi v času sinteze izberemo najprimernejšo. Na ta način se želi zmanjšati potrebo po uporabi postopkov za manipulacijo z govornimi signali, kot je npr. PSOLA (Moulines in Charpentier, 1990), ki znatno poslabšajo kakovost govornega signala.

Poleg korpusne sinteze je na področju sinteze govora opaziti tudi večji poudarek na inženirskih tehnikah (iskalni postopki, optimizacijski postopki, statistično modeliranje), nekoliko manj pa je opaziti razvoja jezikoslovnih pravil (Ostendorf in Bulyko, 2002). Precej inženirskih tehnik izvira iz sicer sorodnega področja razpoznavanja govora, med katerima pa v preteklosti ni bilo dosti pretoka znanj. Kot morda najbolj izstopajoč primer velja omeniti za področje razpoznavanja govora zelo značilno tehnologijo prikritih Markovovih modelov (PMM), ki se v zadnjem času vse pogosteje uporablja za samodejno segmentacijo in označevanje zbirk govorjenega jezika (Mihelič in sod., 2003).

V skladu s trenutnimi trendi predstavljamo na PMM-jih temelječ pristop k sintezi slovenskega govora. Preostanek članka je razdeljen na tri razdelke. V prvem je na kratko predstavljena glavna ideja sinteze govora z uporabo PMM-jev. V drugem podrobneje podamo postopek gradnje sistema in predstavimo poskuse, s katerimi smo ovrednotili kakovost govora. V zadnjem razdelku začrtamo smer, ki ji bomo poskusili slediti v prihodnje.

2. Modeliranje in tvorjenje govornega signala

Postopek sinteze govora z uporabo PMM-jev se razlikuje od bolj razširjenih postopkov v tem, da ogrodje PMM-jev ne uporablja zgolj za segmentacijo in označevanje govorne zbirke, pač pa gre še korak naprej in ga uporablja tudi kot model za tvorjenje govora. Ideja je bila prvič predlagana v (Tokuda in sod., 1995) in kasneje razširjena v (Yoshimura in sod., 1998).

Shematski prikaz sinteze govora z uporabo PMM-jev je prikazan na sliki 1. Na vrhu je uprizorjen *postopek učenja*, pri katerem se ocenijo parametri statističnega modela govora (srednji del slike), na dnu slike pa je uprizorjen *postopek sinteze*, kjer pride do tvorjenja govornega signala.

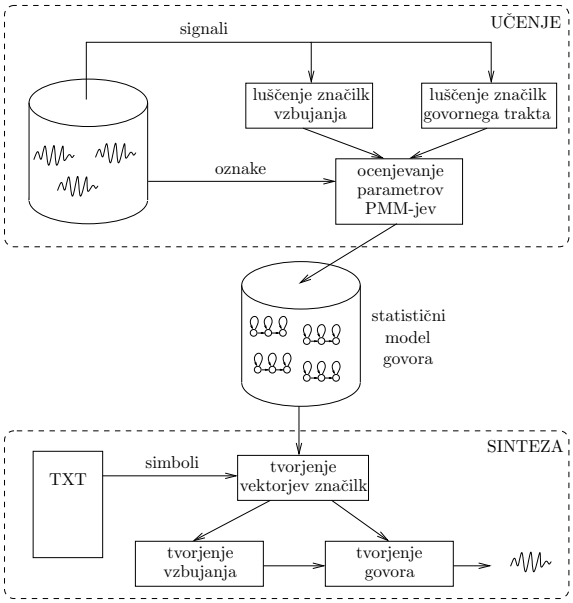
2.1. Parametrizacija govornega signala

Za uspešno ocenjevanje statističnega modela je potrebno zmanjšati entropijo (zgostiti informacijo) izvornega govornega signala. To storimo tako, da govor parametriziramo.

Pri izbiri primernih značilk, ki jih bomo uporabili za sintezo govora, izhajamo iz dejstva, da moramo iz značilk rekonstruirati govorni signal, ki bo v zaznavnem smislu čim bolj podoben izvornemu signalu. V ta namen se izkaže zelo uporabna *teorija vir-filter modela govora* (Rabiner in Huang, 1993). V skladu s to teorijo govor razstavimo na *prenosno funkcijo govornega trakta* in *vzbujanje*, ki ga lahko zadovoljivo opišemo s potekom osnovne frekvence.

2.2. Statistično modeliranje govora

Postopki za ocenjevanje parametrov PMM-jev so dobro znani s področja razpoznavanja govora in jih lahko s pridom uporabimo tudi v naš namen. Kljub temu pa je prisotna pomembna razlika, saj pri gradnji razpoznavalnika govora ocenjujemo le parametre modela prenosne funkcije



Slika 1: Shematski prikaz postopka učenja in sinteze govora z uporabo prikritih Markovovih modelov.

govornega trakta, medtem ko tukaj želimo zgraditi tudi model vzbujanja. Na prvi pogled se zdi, da bi lahko vektor značilk, s katerimi opišemo prenosno funkcijo govornega trakta, enostavno razširili z vrednostjo osnovne frekvence. Vendar naletimo na težavo, saj osnovna frekvenca črpa vrednosti iz množice (pozitivnih) realnih števil le pri zvenečih delih govora. Elegantna rešitev tega problema je ponujena v (Tokuda in sod., 2002), kjer je predlagan nov tip PMM-ja. Pravimo mu PMM z večprostorskimi porazdelitvami (ang. Multi Space Distribution HMM, MSD-HMM) in predstavlja posplošitev tako zveznega kot tudi diskretnega PMM-ja, saj oba vključuje kot poseben primer.

Poleg prenosne funkcije in vzbujanja govornega trakta želimo s PMM-ji modelirati tudi trajanja glasov v govoru. Kljub temu, da PMM-ji preko *matrike prehodnih verjetnosti* implicitno že vsebujejo informacijo o trajanjih stanj, se izkaže, da je eksponentna funkcija gostote verjetnosti neprimerna za večino fizikalnih signalov (Rabiner in Huang, 1993). Težavo lahko rešimo tako, da v PMM vključimo eksplicitne gostote verjetnosti (parametrične ali neparametrične) trajanj stanj (Ferguson, 1980; Levinson, 1986). Vendar s tem znatno povečamo časovno zapletenost postopkov učenja, hkrati pa potrebujemo za dovolj natančno ocenitev parametrov PMM-jev še več učnih podatkov. Problem poenostavimo tako, da ocenimo funkcije verjetnosti trajanj stanj po že končanem postopku učenja (Yoshimura in sod., 1998).

3. Opis poizkusov in vrednotenje kakovosti sintetiziranega govora

V naslednjih podrazdelkih bosta najprej podrobneje opisana postopka učenja in sinteze, nato bodo predstavljeni še testi, s katerimi smo ovrednotili kakovost sintetiziranega govora.

3.1. Postopek učenja

Pri delu smo uporabljali del govorne zbirke vremenskih napovedi VNTV (Žibert in Mihelič, 2000), ki ga je izgovoril govorec `02m`. V zbirki je govorec `02m` zastopan s 578 stavki, 6363 (770 različnih) besedami oz. 39 minutami govora.

Pri izbiri osnovnih govornih enot smo izhajali iz (Zemljak in sod., 2000). V nasprotju z uveljavljeno prakso pri razpoznavanju govora, smo se odločili, da bomo ločevali med dolgimi in kratkimi samoglasniki, saj želimo v sintetiziranem govoru ohraniti čim več prozodičnih značilnosti iz naravnega govora. Izmed vseh alofonskih različic nekega fonema smo ohranili le tiste, ki se morejo različno izgovarjati pri istem levem in desnem kontekstu, skupaj 38.

Značilke smo določali na 25 ms dolgih izsekih govornega signala, ki smo jih predhodno oknili z Blackmanovim oknom. Zaporedni izseki so si sledili s 5 ms zamikom. Prenosno funkcijo govornega trakta smo opisali s 25 koeficienti MFCC s pripadajočimi dinamičnimi (Δ in $\Delta\Delta$) parametri. Vzbujanje smo opisali z vrednostjo $\log f_0$ in prav tako dodanima Δ in $\Delta\Delta$ parametroma — skupaj 78 razsežen vektor značilk.

Izbrali smo levo-desno topologijo PMM-jev brez preskokov. Značilke smo razdelili na štiri neodvisne tokove (prvi za MFCC-je s pripadajočimi dinamičnimi značilkami, drugi za $\log f_0$, tretji za $\Delta \log f_0$ in četrti za $\Delta\Delta \log f_0$). Značilke MFCC smo izračunali s pomočjo zbirke orodij SPTK, poteke osnovne frekvence pa smo poiskali z orodjem `getf0` (Talkin, 1995) iz programskega paketa ESPS. Vsa stanja PMM-jev so vsebovala po eno 75-razsežno Gaussovo funkcijo gostote verjetnosti (prvi tok) in tri dvoprostorske porazdelitve (drugi, tretji in četrti tok). Prvi podprostor je vseboval enorazsežno Gaussovo funkcijo gostote verjetnosti, drugi podprostor pa je bil brezdimenzijski. Vse normalne gostote verjetnosti smo opisali s povprečnim vektorjem in diagonalno kovariančno matriko.

Za učenje smo uporabljali različico orodja HTK, ki dovoljuje uporabo PMM-jev z večprostorskimi porazdelitvami (Tokuda in sod., 2002). Posnetke govornih signalov smo najprej časovno poravnali s pomočjo samodejnega postopka *siljenega prileganja* (Dobrišek, 2001). Nato smo izvedli en prehod Viterbijevega postopka učenja, s katerim smo dobili začetne ocene parametrov modelov. Sledilo je 10 prehodov Baum-Welchevega postopka učenja, s čimer smo dobili natančnejše ocene.

Vse videne trifone smo tvorili kot kopije ustreznih monofonov, nakar smo izvedli še nekaj prehodov učnega postopka. Da bi zmanjšali število parametrov, smo izvedli *vezavo parametrov* na podlagi *fonetičnih vprašanj* (Dobrišek, 2001). Temu je sledilo spet nekaj iteracij učnega postopka, s čimer smo dobili končne ocene.

Po koncu postopka učenja smo na podlagi statistik, ki smo jih pridobili med učenjem, ocenili še parametre *modela trajanj*, ki je vseboval eno enorazsežno funkcijo gostote verjetnosti na stanje PMM-ja.

3.2. Postopek sinteze

Na podlagi vhodnega niza simbolov se ustrezni PMM-ji povežejo v verigo (kompozitni model λ). Naša naloga je poiskati najverjetnejši niz vektorjev značilk $\hat{x}$, ki ga

model λ odda. Z drugimi besedami, poiskati želimo niz $\hat{x} = \{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_T\} = \arg\max_{\mathcal{X}} P(x|\lambda)$. Čeprav je bil za rešitev tega problema predlagan EM tip postopka (Tokuda in sod., 2000), je s praktičnega vidika bolj sprejemljiva podoptimalna rešitev, ki jo dobimo tako, da poenostavimo in najprej najdemo najverjetnejšo pot $\hat{q} = \{\hat{q}_1, \ldots, \hat{q}_T\} = \arg\max_{\mathcal{Q}} P(q|\lambda)$ skozi model λ in nato še najverjetnejši niz $\hat{x} = \arg\max_{\mathcal{X}} P(x|\hat{q}, \lambda)$ na tej poti.

Najverjetnejšo pot najdemo na podlagi modela trajanj, ki smo ga zgradili v postopku učenja. Za dan čas T želimo poiskati niz stanj $\hat{q} = q_1 q_2 \ldots q_T$, za katerega bo verjetnost $P(\hat{q}|\lambda)$ največja. Veljati mora:

$$\frac{\partial}{\partial d_i} \log P(q|\lambda) = \sum_{i=1}^{L} \log p_i(d_i) \stackrel{!}{=} 0, \qquad (1)$$

za $i = 1, 2, \ldots, L$ ob pogoju $\sum_{i=1}^{L} d_i = T$. Če upoštevamo dejstvo, da funkcije gostote verjetnosti $p_i(d_i)$ modeliramo z enorazsežnimi Gaussovimi funkcijami, je rešitev gornjega sistema enačb za $i = 1, 2, \ldots, L$ sledeča:

$$d_i = \mu_i + \frac{T - \sum_{k=1}^{L} \mu_k}{\sum_{k=1}^{L} \sigma_k^2} \sigma_i^2 = \mu_i + \rho \sigma_i^2, \qquad (2)$$

Vidimo, da moremo hitrost govora kontrolirati na dva načina; relativno, z nastavljanjem parametra ρ, in absolutno, z nastavljanjem skupnega časa T.

Podobno poiščemo tudi najverjetnejši niz $\hat{x}$, ki ga model λ odda na poti $\hat{q}$, določenimi z vrednostmi d_i. Če odvajamo funkcijo $\log P(x|\hat{q}, \lambda)$ po vseh spremenljivkah $\mathbf{x}_i$, kjer gre i od 1 do T, dobimo sistem linearnih enačb. Da se pokazati (Tokuda in sod., 2000), da je v primeru, ko imamo samo statične značilke, rešitev tega sistema preprosto niz povprečnih vektorjev. Po drugi strani pa upoštevanje omejitev, ki jih vnesejo dinamične značilke, vodi k bolj realističnim rešitvam. Zainteresirani bralec naj se za detajlni opis postopkov določanja najbolj verjetnega niza značilk obrne na (Tokuda in sod., 2000).

Iz dobljenega niza parametrov (MFCC-jev in vrednosti $\log f_0$) lahko tvorimo govorni signal direktno z uporabo MLSA filtra (Fukada in sod., 1992).

3.3. Vrednotenje

Vrednotenje kakovosti sintetiziranega govora smo razdelili na dva dela. V prvem delu smo izvedli slušni preizkus, v drugem pa smo skušali ugotovitve slušnega preizkusa podpreti še z objektivnimi postopki.

3.3.1. Slušni preizkus

Slušni test smo zastavili v obliki *primerjalnega testa parov* (Gros in sod., 1997). Med seboj smo primerjali štiri različice na PMM-jih temelječe sinteze govora:

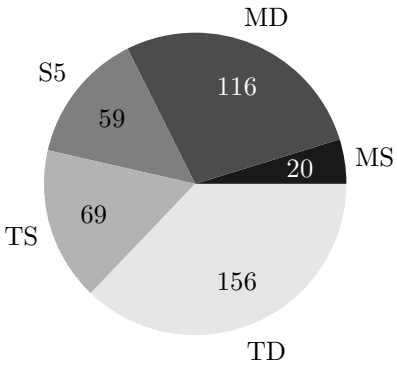
- monofonski modeli, statične značilke (MS),
- monofonski modeli, dinamične značilke (MD),
- trifonski modeli, statične značilke (TS),
- trifonski modeli, dinamične značilke (TD).

Vse štiri različice smo učili v skladu s postopkom, ki smo ga opisali v prejšnjem poglavju, le da smo pri monofonskih različicah (MS in MD) postopek ustavili takoj po delu postopka, ki se nanaša na učenje monofonskih modelov. Nadaljna razlika je bila ta, da smo v primeru statičnih različic (MS in TS) v postopku sinteze parametre tvorili brez upoštevanja omejitev, ki jih vnesejo zveze med statičnimi in dinamičnimi značilkami, medtem ko smo v primeru dinamičnih različic (MD in TD) te omejitve upoštevali. Na ta način smo želeli preveriti vpliv upoštevanja (trifonskega) konteksta in dinamičnih značilk na kakovost sintetiziranega govora.

Da bi sintezo, ki temelji na PMM-jih primerjali tudi z difonsko sintezo, smo v eksperiment vključili še difonski sintetizator (S5) (Gros in sod., 1997).

Posnetke sintetiziranega govora smo tvorili iz treh sestavkov osmih povedi. Prvi dve besedili nista bili povezani s področjem vremenskih napovedi, tretje pa je bilo del vremenske napovedi, ki pa ni bila zajeta v učni množici. Na ta način smo se želeli izogniti neposredni prednosti sintetizatorjev, ki so bili učeni iz govorne zbirke vremenskih napovedi, pred difonskim sintetizatorjem. Iz vsakega sestavka smo tvorili pet različnih sintetiziranih posnetkov in med njimi v naključnem vrstnem redu tvorili vseh deset dvojic. Poslušalec se je ob poslušanju, glede na splošen vtis, pri vsaki dvojici odločil za en posnetek.

V preizkusih je sodelovalo 14 oseb starih med 20 in 30 let, od tega 10 žensk in 4 moški. Nihče na področju govornih tehnologij ni imel strokovnih izkušenj. Rezultati (slika 2) navajajo na naslednje sklepe:



Slika 2: Rezultati slušnih preizkusov.

- h kakovosti sintetiziranega govora največ doprinese postopek tvorjenja parametrov govornega signala, ki upošteva dinamične značilke,
- dodatno izboljšanje kakovosti dosežemo z vpeljavo kontekstno odvisnih (trifonskih) modelov,
- splošni vtis sinteze TS je na ravni difonske S5.

Kot zanimivost omenimo, da ni bilo mogoče opaziti bistvenih razlik v ocenah poslušalcev glede na tematiko besedila, kot bi bilo mogoče pričakovati. Za zaneslivejšo primerjavo bi bilo verjetno potrebno zastaviti nov preizkus, v katerem bi namesto glede na samo tematiko besedila opazovali razlike v ocenah poslušalcev glede na videnost sintetiziranih glasovnih enot v učnem gradivu.

3.3.2. Objektivni preizkusi

Da bi preverili verodostojnost gornjih rezultatov, smo izvedli še nekaj objektivnih testov.

Izhajali smo iz označenih posnetkov izgovorjenega besedila, sestavljenega iz 34 stavkov, ki jih je izgovoril govorec `02m`, a niso bili vsebovani v učni množici. Iz fonemskega zapisa besedila smo tvorili ustrezne signale sintetiziranega govora z različnimi načini sinteze govora.

Iz posnetkov smo izračunali nize vektorjev MFCC, med njimi pa s postopkom *ukrivljanja časovne osi* (Rabiner in Huang, 1993) normirane Evklidove razdalje po sledeči formuli:

$$d(\tilde{c}^{(1)}, \tilde{c}^{(2)}) = \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{k=1}^{M-1} \left(\tilde{\mathbf{c}}^{(1)}_{w_1(t)}(k) - \tilde{\mathbf{c}}^{(2)}_{w_2(t)}(k) \right)^2 \right)^{\frac{1}{2}}, \tag{3}$$

kjer sta $w_1(t)$ in $w_2(t)$ funkciji ukrivljanja časovne osi, ki pripadata nizoma $\tilde{c}^{(1)}$ in $\tilde{c}^{(2)}$ $(M-1)$-razsežnih vektorjev značilk.

Ugotovili smo (Vesnicer, 2003) sledeče:

- razdalje med sintetiziranim govorom in naravnim govorom ($> 1{,}4$) so precej večje kot razdalje med različnimi verzijami sintetiziranega govora ($< 0{,}4$),
- paroma je najmanjša razdalja med monofonskima sintetizatorjema ($\approx 0{,}2$) in trifonskima sintetizatorjema ($\approx 0{,}2$).

Preveriti smo želeli tudi razlike v trajanju med naravnim in sintetiziranim govorom. Izračunali smo povprečna odstopanja d_i trajanj istoležnih glasov,

$$d_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \left| t^{(S)}_{i,j} - t^{(N)}_{i,j} \right|, \quad i = 1, 2, \ldots, 38, \tag{4}$$

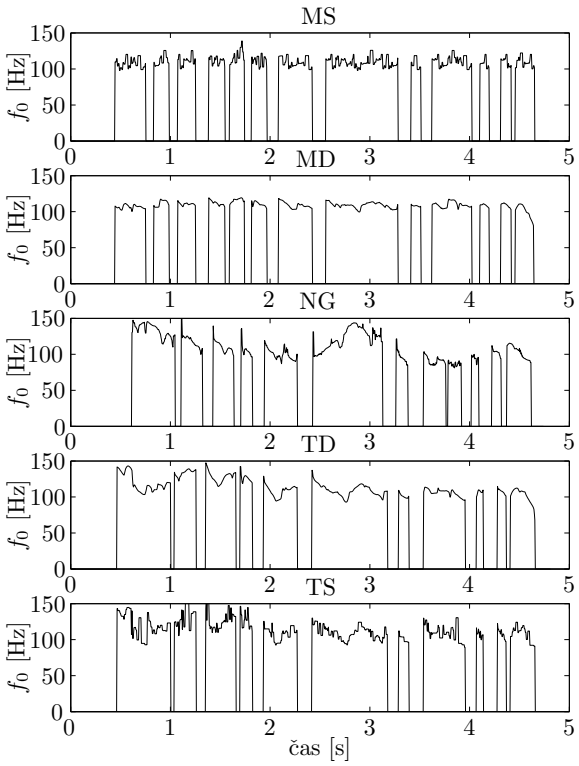
kjer je $t^{(S)}_{i,j}$ trajanje j-te pojavitve i-tega glasu sintetiziranega govora, $t^{(N)}_{i,j}$ trajanje j-te pojavitve i-tega glasu naravnega govora, N_i pa število pojavitev i-tega glasu, ter povprečno odstopanje d,

$$d = \frac{1}{N} \sum_{i=1}^{38} \sum_{j=1}^{N_i} \left| t^{(S)}_{i,j} - t^{(N)}_{i,j} \right|, \tag{5}$$

kjer je $N = \sum_{i=1}^{38} N_i$. Pri monofonski različici znaša d 20 ms, pri trifonski pa 17 ms. Dodatno opazimo (Vesnicer, 2003), da je povprečno odstopanje trajanj glasov trifonske različice sintetizatorja pri večini glasov manjše kot pri monofonski različici. Čeprav rezultati niso direktno primerljivi, povejmo, da je z dvostopenjskim modelom trajanja za slovenski govor (Gros, 1999) dosežena vrednost povprečne absolutne razlike trajanj 11 ms.

Za konkreten primer izgovorjenega stavka "*V gorah bo ponoči in zjutraj še oblačno in megleno, občasno bo rahlo snežilo.*" smo določili poteke f_0 (slika 3) pri naravnem govoru (NG) in štirih prej omenjenih različicah sintetiziranega govora. Razberemo lahko dvoje:

- pri "statičnih" različicah sinteze govora (MS in TS) prihaja do nezveznosti v poteku osnovne frekvence, medtem ko je pri "dinamičnih" različicah MD in TD potek osnovne frekvence zglajen,



Slika 3: Primerjava potekov osnovne frekvence.

- potek osnovne frekvence je pri monofonskih različicah MS in MD precej monoton, medtem ko je pri trifonskih TS in TD bolj razgiban.

Strniti moremo, da dinamične značilke prispevajo k bolj gladkim prehodom med glasovi v sintetiziranem govoru, kar prispeva predvsem k večji razločnosti in razumljivosti govora, kontekstno odvisni modeli (trifoni) pa prispevajo k večji razgibanosti govora in s tem k bolj naravnemu govoru.

4. Sklep

Predstavili smo sistem za sintezo govora, ki temelji na PMM-jih. Kakovost sintetiziranega govora smo ovrednotili s slušnim preizkusom in različnimi objektivnimi preizkusi. Ti potrjujejo, da so številne značilnosti naravnega govora ohranjene tudi v sintetiziranem govoru. Razumljivost in naravnost govora se po pričakovanju znatno poveča z uporabo dinamičnih značilk in kontekstno odvisnih modelov.

V prihodnje se bomo posvetili problemu, kako v modelu govora zaobjeti še več prozodične informacije, ki v veliki meri prispeva k naravnosti govora.

5. Literatura

N. Campbell in A. Black, 1996. *Prosody and the Selection of Source Units for Concatenative Synthesis*, str. 279–282.

S. Dobrišek. 2001. *Analysis and Recognition of Phones in Speech*. Ph.d. diss. (in slovene), University of Ljubljana.

J. D. Ferguson. 1980. Variable duration models for speech. V: *Proc. of the symposium on the Application of Hidden Markov Models to Text and Speech*, str. 143–179.

T. Fukada, K. Tokuda, T. Kobayashi in S. Imai. 1992. An adaptive algorithm for mel-cepstral analysis of speech. V: *Proc. ICASSP*, str. 137–140.

J. Gros, N. Pavešić in F. Mihelič. 1997. Text-to-speech synthesis: A complete system for the slovenian language. *CIT*, 5(1):11–19.

J. Gros. 1999. Dvostopenjski model trajanja za slovenski jezik. *Elektrotehniški vestnik*, 66(2):92–97.

S. E. Levinson. 1986. Continuously variable duration hidden markov models for automatic speech recognition. *Computer, Speech and Language*, 1(1):29–45.

F. Mihelič, J. Gros, S. Dobrišek, J. Žibert in N. Pavešić. 2003. Spoken language resources at luks of the university of ljubljana. *International Journal of Speech Technology*, 6:221–232.

E. Moulines in F. Charpentier. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–468.

M. Ostendorf in I. Bulyko. 2002. The impact of speech recognition on speech synthesis. V: *Proc. IEEE Workshop on Speech Synthesis*.

L. Rabiner in B.-H. Huang. 1993. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ, USA.

D. Talkin, 1995. *A Robust Algorithm for Pitch Tracking (RAPT)*, str. 495–518.

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

K. Tokuda, T. Yoshimura in T. Masuko. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. V: *Proc. ICASSP*, zvezek 3, str. 1315–1318.

K. Tokuda, T. Masuko in N. Miyazaki. 2002. Multi-space probability distribution HMM. *IEICE Transactions on Information and Systems*, E85-D(3):455–464.

B. Vesnicer. 2003. *Umetno tvorjenje govora z uporabo prikritih Markovovih modelov*. Magistrska nalog, Fakulteta za elektrotehniko, Univerza v Ljubljani.

J. Žibert in F. Mihelič. 2000. Slovenian weather forecast speech database. V: *Proc. SoftCOM*, zvezek 1, str. 199–206.

T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi in T. Kitamura. 1998. Duration modeling for HMM-based speech synthesis. V: *Proc. ICSLP*, zvezek 2, str. 29–32.

M. Zemljak, Z. Kačič, S. Dobrišek, J. Gros in P. Weiss. 2000. Computer-based symbols for slovene speech. *Journal for Linguistics and Literary Studies*, 2:159–294.