Sinteza govora z uporabo prikritih Markovovih modelov

Boštjan Vesnicer, France Mihelič, Nikola Pavešić

Fakulteta za elektrotehniko, Univerza v Ljubljani
Tržaška 25, 1001 Ljubljana, Slovenija
{bostjan.vesnicer,france.mihelic,nikola.pavesic}@fe.uni-lj.si

Povzetek

V tem članku je predstavljen postopek sinteze govora s pomočjo prikritih Markovovih modelov. Osnovo za pridobivanje značilk, s katerimi učimo modele, predstavlja vir–filter model govora. Vektor značilk tako sestavimo iz dveh delov. Prvi del opisuje govorni trakt, drugi del pa vzbujanje. Vhod v postopek sinteze predstavlja niz fonemov s pripadajočimi trajanji in poteki osnovne frekvence. Govor nato tvorimo tako, da iz ustreznih prikritih Markovovih modelov generiramo najverjetnejši niz vektorjev značilk, iz katerega nato rekonstruiramo govor. Če vektor značilk razširimo še z dinamičnimi značilkami, dobimo gladke prehode med posameznimi glasovi.

1. Uvod

Čeprav so sistemi za sintezo govora (angl. text-to-speech systems) že dosegli visoko stopnjo razumljivosti sintetiziranega govora, jim še večjo uveljavitev v vsakdanjih aplikacijah preprečuje predvsem to, da je sintetiziran govor še vedno precej nenaraven.

Za sintezo[1] govora obstajajo tri glavne metode. To so artikulatorna sinteza, formantna sinteza ter sinteza z združevanjem (angl. concatenative synthesis). Pri *artikulatorni* sintezi skušamo modelirati statične in dinamične lastnosti človeških govornih organov (npr. glasilk, jezika, žrela, ustne votline in ustnic). *Formantno sintezo* lahko v splošnem obravnavamo kot spisek pravil (zato govorimo o t.i. sintezi na podlagi pravil (angl. rule-based synthesis)), s katerimi opišemo resonančne frekvence govornega trakta (*formante*) in iz katerih tvorimo govor. Taki sistemi dajejo razmeroma razumljiv, a hkrati precej nenaraven govor. To je tudi razumljivo, saj je težko tako kompleksen proces kot je tvorjenje govora natančno zajeti z zbirko enostavnih pravil. Večjo naravnost lahko dosegamo z drugim načinom sinteze, ki ji drugače pravimo tudi sinteza na podlagi podatkov (angl. data-based synthesis). Pri takem postopku je najprej potrebno posneti osnovne govorne enote (npr. difone) in jih tudi primerno označiti. Te enote med samo sintezo združujemo v govor s posebnim postopkom, najpogosteje je to PSOLA (angl. Pitch-Synchronous-Overlap-and-Add), ki hkrati služi za oblikovanje intonacije govora. Kljub temu, da na tak način sintetiziran govor dosega dokaj visoko kvaliteto (ta se meri s posebnimi slušnimi testi), je največja ovira še vedno nezadostna naravnost. V zadnjem času se korak k večji naravnosti govora poskuša storiti s posplošitvijo tega postopka tako, da sistem primerne govorne enote izbira dinamično (v času sinteze) iz večje govorne zbirke. Lastnost takšnih enot je tudi ta, da njihova dolžina ni v naprej predpisana. Takemu načinu sinteze govora pravimo sinteza z uporabo velike govorne zbirke (angl. corpus-based synthesis) ali z drugim imenom sinteza z izbiro enot (angl. unit-selection synthesis). Največji problem tega postopka predstavlja izbira kriterijske funkcije, s pomočjo katere se sistem odloča o primernosti posameznih enot.

Med postopke sinteze govora s pomočjo pravil lahko štejemo tudi postopek, ki se opira na vir–filter model govora. Za razliko od prejšnjega postopka skušamo v tem primeru govor zajeti v parametrični obliki. S filtrom želimo modelirati prenosno funkcijo govornega trakta, z virom pa vzbujanje. Prednosti parametričnega modela govora sta dve. S parametričnim opisom dosežemo v smislu zgoščevanja informacije bolj kompaktno predstavitev govorne zbirke in hkrati večjo fleksibilnost v smislu spreminjanja prozodičnih lastnosti govora (trajanje in potek osnovne frekvence).

V tem delu želimo izkoristiti dobre lastnosti vir–filter modela govora in hkrati doseči večjo naravnost sintetiziranega govora. Temu cilju se želimo približati s pomočjo statističnega modela govora. Enako kot pri razpoznavanju govora želimo v fazi učenja govor modelirati s pomočjo prikritih Markovovih modelov (PMM), razlika pa je v tem, da želimo sedaj iz naučenih modelov znati generirati govor.

V drugem delu je predstavljen vir–filter model govora, ki se pogosto uporablja pri kodiranju govora, nam pa bo služil za generiranje značilk, s katerimi bomo učili PMM-je. V tretjem delu je podan postopek generiranja parametrov govornega signala iz naučenih PMM-jev. V četrtem delu so opisane značilnosti postopka sinteze govora z uporabo PMM-jev. V petem delu je opisan poskus sinteze govora, v zaključku pa so podani še načrti za prihodnje delo.

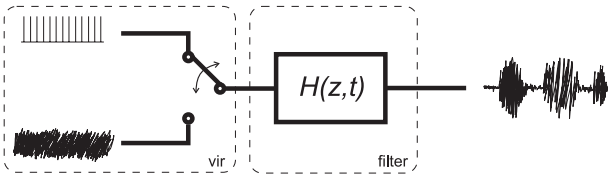
2. Vir–filter model govora

Osnovna ideja pri vir–filter modelu govora je, da želimo govorni signal ločeno opisati s časovno spremenljivo prenosno funkcijo govornega trakta $H(z,t)$ in pa z vzbujanjem, ki ta govorni signal generira (slika 1). Prenosno funkcijo govornega trakta lahko opišemo na več načinov. Eden izmed bolj znanih postopkov pri reševanju tega problema je postopek linearnega napovedovanja (angl. linear prediction), ki se zelo pogosto uporablja pri kodiranju govornega signala.

2.1. Predstavitev govornega trakta s koeficienti linearnega napovedovanja

Pri linearni predikciji predpostavimo, da je vrednost signala v določenem trenutku linearna kombinacija

[1] V okviru tega prispevka s *sintezo* mislimo na generiranje govornega signala (angl. waveform generation), kar predstavlja le enega izmed problemov pri sintezi govora iz teksta.



Slika 1: Vir–filter model govora. Stikalo preklaplja med generatorjem šuma (nezveneči glasovi) in generatorjem impulzov (zveneči glasovi).

določenega števila preteklih vrednosti.

$$s(n+1) = \sum_{k=0}^{k=N} a_k s(n-k), \tag{1}$$

kjer koeficiente a_k imenujemo koeficienti linearne predikcije (LPC) in predstavljajo formante govornega trakta. Ko izračunamo te koeficiente, lahko vzbujanje govornega trakta poiščemo z inverznim filtriranjem govornega signala. Izkaže se, da lahko poenostavljeno vzbujanje za zveneče dele govora aproksimiramo kar z vlakom impulzov, katerih perioda ustreza osnovni periodi govora, ter z belim šumom za nezveneče dele govora. Vzbujanje bi tako lahko opisali le z enim parametrom, ki bi predstavljal osnovno frekvenco govora, na nezvenečih delih pa bi vrednost tega parametra postavili kar na nič. Takšno parametrizacijo vzbujanja dosežemo torej z detektorjem osnovne periode govora.

Iz dobljenih parametrov za vzbujanje in prenosno funkcijo govornega trakta bi lahko spet enostavno tvorili govorni signal tako, da bi v odvisnosti od parametra, ki predstavlja osnovno frekvenco vzbujanja, preklapljali med šumnim generatorjem in generatorjem vlakov impulzov, kar bi predstavljalo vzbujanje za filter, ki smo ga opisali z LPC koeficienti.

Očitno je, da z opisanim modelom vzbujanja vnesemo precejšnjo poenostavitev. Določeni glasovi nastanejo namreč s kombinacijo obeh načinov vzbujanja. To dejstvo poskuša upoštevati mešani model vzbujanja (angl. mixed excitation), ki je uporabljen pri MELP (McCree et al., 1996) različici kodiranja govora. Osnovna ideja je ta, da vzbujanje ločimo na posamezne frekvenčne pasove in za vsak pas posebej določimo, ali je zveneč ali nezveneč. Pri generiranju govora pa na podlagi te analize ločeno spustimo obe vrsti vzbujanja skozi ustrezne pasovne filtre, izhoda iz teh filtrov pa seštejemo. Tako v primeru, da imamo govorni signal vzorčen s 16 kHz, frekvenčno področje razdelimo na pet delov (0–1000, 1000–2000, 4000–6000 in 6000–8000 Hz), za katere jakost zvenečnosti določimo z uporabo normaliziranih korelacijskih koeficientov v okolici osnovne periode. Korelacijski koeficient pri zamiku t je definiran kot

$$c(t) = \frac{\sum_{n=0}^{N-1} s(n)s(n+t)}{\sqrt{\sum_{n=0}^{N-1} s(n)s(n) \sum_{n=0}^{N-1} s(n+t)s(n+t)}}, \tag{2}$$

kjer je $s(n)$ vrednost n–tega vzorca govornega signala, N pa dolžina okna. Dodatno se ocenijo še vrednosti amplitudnega spektra pri prvih deset večkratnikih osnovne periode.

2.2. Predstavitev govornega trakta s koeficienti melodičnega kepstra

Drugačen način opisa govornega trakta dosežemo s koeficienti melodičnega kepstra (MFCC), ki se zelo pogosto uporabljajo v namene razpoznavanja govora, kjer je potrebno za uspešno učenje razpoznavalnikov iz govora izluščiti le tisto informacijo, ki prispeva k večji zanesljivosti razpoznavanja.

Tudi tukaj nas zanima, kako lahko iz MFCC koeficientov rekonstruiramo prvotni govorni signal. V tem primeru kot filter uporabimo MLSA (Mel Log Spectral Approximation) filter,

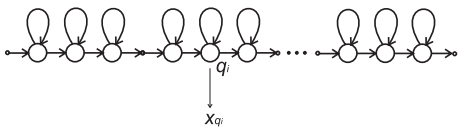
$$H(z) = \exp \sum_{m=0}^{M} c(m) \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \tag{3}$$

kjer so $c(m)$, $0 \leq m \leq M$, koeficienti melodičnega kepstra, parameter α, $|a| < 1$, pa podaja ukrivljenost melodične skale.

3. Postopek tvorjenja parametrov govornega signala iz prikritih Markovovih modelov

Pri razpoznavanju govora želimo na podlagi vhodnega niza vektorjev značilk poiskati najbolj verjetno pot skozi graf PMM-jev, kar storimo z Viterbijevim algoritmom. Pri sintezi pa prejmemo na vhodu niz simbolov, ki ustrezajo posameznim modelom. Na podlagi tega niza simbolov želimo poiskati najbolj verjeten niz značilk, $\hat{x}$, ki ga lahko veriga, zgrajena iz ustreznih osnovnih modelov, generira.

Ilustrirajmo problem še na konkretnem primeru. Denimo, da želimo sintetizirati besedo *primer*, kar pomeni, da opisani postopek prejme niz alofonov *p,r,i,m,e:,r*. Na podlagi tega niza povežemo osnovne PMM modele posameznih alofonov v verigo in dobimo sestavljeni PMM model λ, iz katerega želimo poiskati najbolj verjeten niz vektorjev značilk $\hat{x}$, ki ga lahko ta model generira.



Slika 2: Sestavljen PMM λ, iz katerega generiramo najverjetnejši niz značilk $\hat{x}$.

Na sliki 2 smo s q_i označili stanje, v katerem se PMM λ nahaja v trenutku t_i, z $\boldsymbol{x}_{q_i}$ pa vektor značilk, ki ga odda PMM λ pri prehodu iz stanja q_{i-1} v stanje q_i.

Povzemimo metodo, ki je bila objavljena v (Tokuda et al., 1995).

Naj bo λ levo-desni PMM (veriga), zgrajen iz N stanj, ki pri prehodu iz stanja q_{i-1} v stanje q_i odda M–razsežen vektor značilk $\boldsymbol{x}_{q_i}$,

$$\boldsymbol{x}_{q_i} = \left(x_1^{(q_i)}, x_2^{(q_i)}, \ldots, x_M^{(q_i)}\right)^T. \tag{4}$$

Sedaj želimo iz modela λ generirati tak niz vektorjev značilk $\hat{x}$, $\hat{x} = \boldsymbol{x}_{q_1}\boldsymbol{x}_{q_2}\ldots\boldsymbol{x}_{q_L}$, dolžine L, da bo

$$\begin{aligned}\hat{x} &= \arg\max_x \{P(x|\lambda, L)\} \\ &= \arg\max_x \left\{ \sum_Q P(x|q,\lambda)P(q,\lambda) \right\}, \end{aligned} \quad (5)$$

kjer je q, $q = q_1 q_2 \ldots q_L$, pot skozi model λ.

Teoretično je potrebno za določitev niza $\hat{x}$ preiskati vse možne poti q skozi model λ, kar je časovno zelo zahtevno, zato se v praksi poslužimo Viterbijeve aproksimacije

$$\hat{x} \approx \arg\max_x \{P(x|q,\lambda,L)P(q|\lambda,L)\} \quad (6)$$

.

Podoptimalno rešitev lahko poiščemo tako, da najbolj verjetno pot $\hat{q}$ določimo na podlagi informacije o željeni prozodiji sintetiziranega govora, ali pa jo izračunamo neodvisno od niza vektorjev značilk x,

$$\hat{q} = \arg\max_q \{P(q|\lambda, L)\}. \quad (7)$$

Predpostavimo, da izhodno funkcijo gostote verjetnosti i-tega stanja modeliramo z eno Gaussovo funkcijo s povprečnim vektorjem $\boldsymbol{m}_i$ in kovariančno matriko $\boldsymbol{K}_i$. Potem lahko zapišemo

$$\begin{aligned}\ln P(x|q,\lambda) = &-\frac{LM}{2}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{L}\ln|\boldsymbol{K}_{q_t}| \\ &-\frac{1}{2}\sum_{t=1}^{L}(\boldsymbol{x}_{q_t}-\boldsymbol{m}_{q_t})^T\boldsymbol{K}_{q_t}^{-1}(\boldsymbol{x}_{q_t}-\boldsymbol{m}_{q_t}). \end{aligned} \quad (8)$$

Maksimiranje gornjega izraza privede do trivialne rešitve, $\hat{x} = (\boldsymbol{m}_{q_1}, \boldsymbol{m}_{q_2}, \ldots, \boldsymbol{m}_{q_L})$. V primeru statičnih značilk je najbolj verjeten niz kar niz povprečnih vektorjev. Prehodi med posameznimi vektorji pa predstavljajo nezveznosti, tako da na tak način dobimo precej nezvezen govor.

Zanima nas, kaj pridobimo, če statičnim značilkam dodamo še dinamične,

$$\boldsymbol{x}_{q_i} = \left((x_{q_i})^T, (\Delta x_{q_i})^T, \left(\Delta^2 x_{q_i}\right)^T\right)^T. \quad (9)$$

Maksimum izraza 8 sedaj poiščemo tako, da postavimo

$$\frac{\partial}{\partial x_i}\ln P(x|q,\lambda) \equiv \mathbf{0}, \qquad i = 1, 2, \ldots, M, \quad (10)$$

kar prevedemo na reševanje sistema linearnih enačb, ki v splošnem za rešitev zahteva $\mathcal{O}(T^3M^3)$ operacij. V (Tokuda et al., 1995) je podan hiter algoritem za izračun najverjetnejšega niza vektorjev značilk, ki zahteva v najslabšem primeru $\mathcal{O}(T^2M^3)$ operacij.

4. Učenje prikritih Markovovih modelov

Kot govorno zbirko smo uporabili obstoječo zbirko vremenskih napovedi VNTV, ki se že uspešno uporablja pri avtomatskem podnaslavljanju vremenskih napovedi (Žibert et al., 2000; Žibert and Mihelič, 2000). Zbirka obsega pet različnih govorcev, od tega so štirje govorci in ena govorka. V zbirko je zajetih 3882 stavkov, ki so sestavljeni iz 2857 različnih besed, kar znese 252 minut govornega materiala.

Pri razpoznavanju govora se lahko odločimo za od govorca neodvisno ali za od govorca odvisno razpoznavanje. V prvem primeru v učni bazi zajamemo posnetke le enega govorca, v drugem primeru pa želimo zajeti čim večje število različnih govorcev. Obe možnosti imamo na voljo tudi pri sintezi, vendar se zdi bolj smiselno, da se osredotočimo le na tistega govorca, katerega glas želimo uporabiti pri sintezi. Zato smo se odločili, da bomo v prvi fazi za učenje uporabili posnetke le enega govorca. Učna množica je tako obsegala okrog 500 stavkov s skupno dolžino 40 minut.

Učenje prikritih Markovovih modelov smo izvedli z orodjem HTK (Young et al., 2002). Struktura PMM modelov je bila sledeča. Imeli smo 35 monofonskih modelov, od katerih je vsak model imel tri stanja z eno funkcijo normalne gostote verjetnosti na stanje. Dimenzija vektorja značilk je bila 40, od tega smo prenosno funkcijo govornega trakta opisali z 21 koeficienti melodičnega kepstra, ostalih 19 značilk pa smo uporabili za opis vzbujanja. Značilke smo računali na vsakih 5 ms.

5. Značilnosti postopka sinteze govora z uporabo PMM

Opisani postopek sinteze govora s pomočjo prikritih Markovovih modelov nima funkcionalnosti celotnega sistema za pretvorbo teksta v govor (angl. *text-to-speech, TTS*), ampak predstavlja le zadnji modul takega sistema. Predpostavljamo torej, da je nekdo že predhodno opravil analizo vhodnega teksta, grafemsko-fonemsko pretvorbo in oblikoval željeno prozodijo (trajanje in potek osnovne frekvence posameznih fonemov) sintetiziranega govora, ki vsak zase predstavljajo zelo zahtevne probleme (Gros, 2000).

Ena izmed ključnih značilnosti opisanega postopka je tudi ta, da ni od jezika odvisen, če zanemarimo dejstvo, da moramo govorno zbirko pač posneti v jeziku, ki ga želimo sintetizirati. To pomeni, da lahko enak postopek apliciramo na poljuben jezik.

Pomembna prednost pred ostalimi postopki za sintezo govora je še, da je postopek gradnje novega glasu popolnoma avtomatiziran, saj je praktično ekvivalenten učnemu delu postopka pri razpoznavanju govora s PMM-ji. To pomeni, da odpadejo zamudni in zahtevni procesi označevanja govorne zbirke kot jih poznamo npr. pri gradnji difonske zbirke (Gros et al., 1996). Še posebej pa se ta prednost izkaže v primerjavi s korpusno sintezo, kjer imamo opravka z velikimi govornimi zbirkami. V našem primeru gradnja novega glasu ob posedovanju zbirke (lahko tudi enakega glasu v različnih emocionalnih stanjih govorca) namreč s strani načrtovalca zahteva enako dela ne glede na velikost govorne zbirke, če seveda zanemarimo več porabljenega

procesorskega časa. Poleg enostavnosti postopka v fazi učenja pa imamo določeno prednost tudi v fazi sinteze. Modeliranje govora s PMM-ji predstavlja namreč učinkovito kompaktno parametrizacijo govorne zbirke. Velikost modelov, iz katerih tvorimo govor, namreč ni odvisna od velikosti govorne zbirke, iz katere te modele učimo, temveč je odvisna le od strukture modelov, za katero se odločimo. Za primerjavo, zbirke ki se uporabljajo pri korpusni sintezi obsegajo tudi po več ur govornih posnetkov. V času sinteze mora biti prisotna celotna zbirka, ki lahko tako v nekomprimirani obliki zavzame tudi po več sto MB (ena ura govora zavzame pri frekvenci vzorčenja 16 kHz in kvantizaciji 16 bit/vzorec približno 110 MB prostora na disku oz. v pomnilniku). Skupna velikost PMM modelov (35 modelov, 5 stanj na model, 1 izhodna funkcija gostote verjetnosti na stanje, diagonalne kovariančne matrike) pa znaša le okoli 60 kB. S tega stališča se zdi, da so sistemi korpusne sinteze govora mogoče bolj primerni za večje strežniške sisteme, medtem ko bi bili lahko sistemi za sintezo govora, ki temeljijo na prikritih Markovovih modelih, uporabni tudi za vgrajene (angl. embedded) sisteme, ki imajo ponavadi precej omejena sistemska sredstva.

6. Preizkus sintetiziranega govora

Opisan postopek sinteze govora je še v zgodnji fazi razvoja in kot tak še ni primeren za direktno uporabo v kakšni aplikaciji. To je tudi razlog, da resnejših preizkusov kvalitete sintetiziranega govora zaenkrat še nismo opravili. Do sedaj so bili opravljeni le preizkusi, ki so potrdili pravilnost delovanja posameznih postopkov in so služili za grobo oceno kakovosti umetno tvorjenega govora. Ti preizkusi so pokazali, da z opisanim postopkom sinteze ne dosežemo primerljive kakovosti umetnega govora z ostalimi uveljavljenimi postopki. Kljub temu pa smo mnenja, da se da kakovost izboljšati, zato bomo v prihodnje opravili še nadalnje poizkuse.

Vhod v postopek v najboljšem primeru predstavlja niz simbolov, ki ustrezajo posameznim glasovom, s pripadajočimi časi trajanja in potekom osnovne frekvence. Če informacije o času trajanja ali o poteku osnovne frekvence nimamo, lahko postopek še vedno tvori govor z "zglajenim" (v primeru dinamičnih značilk) povprečnim časom trajanja in povprečno osnovno frekvenco posameznih fonemov. Povprečno dolžino trajanja lahko izračunamo na podlagi podatka, kako dolgo ostane PMM v povprečju znotraj določenega stanja, kar izračunamo s sledečo enačbo,

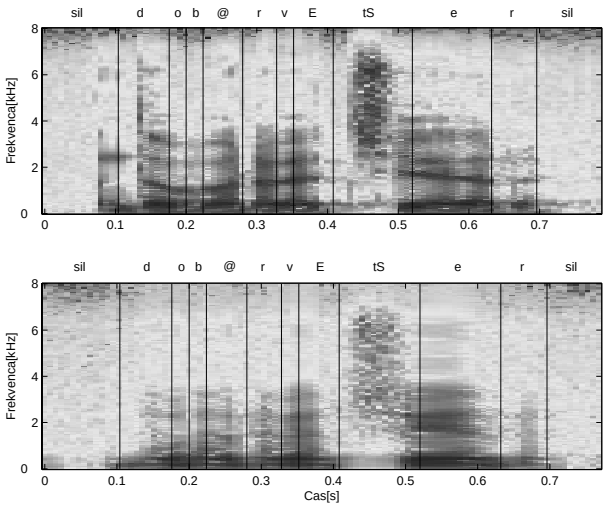
$$E\{d_i\} = \sum_{d=1}^{\infty} d a_{ii}^{d-1}(1 - a_{ii}) = \frac{1}{1 - a_{ii}}, \qquad (11)$$

kjer je a_{ii} verjetnost, da model ostane v stanju i.

Na sliki 3 sta kot primer prikazana spektrograma originalnega posnetka in sintetiziranega govora. Izgovorjen je bil stavek "Dober večer.".

7. Zaključek

Predstavili smo postopek za sintezo govora z uporabo prikritih Markovovih modelov. Osnovo za pridobivanje značilk, ki jih uporabimo za učenje PMM-jev, predstavlja vir–filter model govora.



Slika 3: Spektrogram originalnega posnetka (zgoraj) in sintetiziranega govora (spodaj).

Opravili smo tudi že nekaj začetnih poizkusov, ki kažejo na to, da sintetiziran govor dokaj dobro posnema glas govorca, s katerim smo sistem učili. Z uporabo dinamičnih značilk dosežemo tudi zelo gladke prehode med posameznimi glasovi.

Zaenkrat kvaliteta umetno tvorjenega govora z opisanim postopkom še ni primerljiva z ostalimi uveljavljenimi postopki za sintezo govora.

V prihodnosti nameravamo opraviti še nekaj novih eksperimentov. Izboljšano kakovost sintetiziranega govora bomo poskušali doseči predvsem z vpeljavo kontekstno odvisnih modelov (difoni, trifoni). Pogledali bomo še, kako na kvaliteto sintetiziranega govora vplivajo različne strukture PMM modelov (število stanj na model, model s preskoki, model brez preskokov med stanji, število mešanic funkcij gostot verjetnosti na stanje, diagonalne kovariančne, nediagonalne kovariančne matrike, hitrost računanja značilk). Poleg kepstralnih želimo preizkusiti tudi kakšne druge značilke (npr. LPC). Nenazadnje pa bo potrebno opraviti še tudi bolj obširno evaluacijo kvalitete sintetiziranega govora.

8. Literatura

J. Gros, I. Ipšić, S. Dobrišek, F. Mihelič in N. Pavešić. 1996. Segmentation and labelling of slovenian diphone inventories. *The 16th International Conference on Computational Linguistic, COLING 1996*, 1:298–303.

J. Gros. 2000. *Samodejno tvorjenje govora iz besedil*. Založba ZRC.

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

J. Žibert, F. Mihelič in S. Dobrišek. 2000. Avtomatično podnaslavljanje vremenskih napovedi. *Zbornik devete Elektrotehniške in računalniške konference, ERK 2000*, B:165–168.

A. McCree, K. Truong, E. B. George, T. Barnwell in V. Viswanathan. 1996. A 2.4 kbit/s MELP coder candidate for the new u.s. fedaral standard. *Proc. of the 1996 Intl.*

Conference on Acoustics, Speech, and Signal Processing, 1:200–203.

K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi in S. Imai. 1995. An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features. *Proc. 4th European Conference on Speech Communication and Technology, EUROSPEECH 1995*, 1:757–760.

S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev in P. Woodland. 2002. *The HTK Book*. Cambridge University, 6. izd.