Uspešnost napovedovanja prozodičnih mej z arhitekturo klasifikacije napak samodejno-asociativnih nevronskih mrež

Janez Stergar

Univerza v Mariboru
Fakulteta za elektrotehniko, računalništvo in informatiko
Inštitut za elektroniko
Smetanova ulica 17, 2000 Maribor
janez.stergar@uni-mb.si

Povzetek

Podatkovno vodeni pristopi učenja ponujajo na nekaterih področjih govornih tehnologij – sistemih za samodejno razpoznavanje govora (ASR) in sistemih za samodejno tvorjenje govora iz besedil (TTS) – rešitve v adaptaciji modelov prozodije na različne govorce in specifične zahteve uporabnika (aplikacije). Podatkovno vodeni pristopi omogočajo samodejno izločitev prozodične informacije iz primerne besedilne zbirke naravnega govora. Hkrati pogojujejo takšni pristopi doslednost in konsistentnost pri označevanju besedilnih zbirk, saj je od postopka označevanja velikokrat odvisna uspešnost napovedovanja prozodičnih (in drugih) parametrov iz besedilne zbirke. Prozodija na nivoju besed, besednih zvez, stavkov in povedi – simbolična prozodija – predstavlja pomemben segment v sistemih TTS. Če je napovedovanje simbolične prozodije uspešno, je to pomemben korak k izboljšanju naravnosti sintetiziranega govora. Celovit pristop napovedovanja simbolične prozodije obsega napovedovanje prozodičnih mej, jakostnih izrazitosti v povedi in intonacijske glave povedi. V članku bomo predstavili postopek napovedovanja prozodičnih mej, katerih strukturo smo izločili iz samodejno/ročno označene besedilne zbirke. Preskusili smo uspešnost napovedovanja z arhitekturo klasifikacije napak samodejno-asociativnih nevronskih mrež (AAC). Rezultati uspešnosti napovedovanja izpostavljajo učinkovitost metode označevanja besedilne zbirke, saj smo dosegli dobre rezultate napovedovanja prozodičnih mej že s prepolovljenim obsegom besedilne zbirke (600 povedi). S strukturo AAC smo dosegli uspešnost napovedovanja 89,9 %.

1. Uvod

Sisteme TTS lahko izpostavimo kot odločilno tehnologijo v interaktivnih govornih sistemih in kot nepogrešljiv gradnik v naslednji generaciji sistemov govornega dialoga (Edgington, 1996). Če postopek samodejnega tvorjenja govora iz besedil razčlenimo na dva glavna procesa – *naravno obdelavo besedila*, kjer besedilo prevedemo v ustrezen fonetični prepis, skupaj z želeno intonacijo in ritmom, ter *digitalno obdelavo signalov*, kjer prejeto simbolično informacijo preslikamo v govor (Gros, 1997) – zavzema prozodija pomembnejšo vlogo tako v prvem kot drugem procesu.

V splošnem lahko pojem prozodije razčlenimo na segmentno in nadsegmentno. Formalno bi lahko prozodično strukturo opisali kot hierarhično. Najmanjše enote v tej hierarhiji predstavljajo interne komponente zlogov (npr. glasovi), največje enote pa so segmenti (npr. stavki, povedi, odstavki) (Selkirk, 1984, Hayes, 1995). Pojem simbolične prozodije povezujemo z nadsegmentnimi prozodičnimi lastnostmi. V skladu z jezikoslovno tradicijo se nanaša pojem simbolične prozodije na govorne značilnosti, ki se ne nanašajo na en sam fonetični segment, glas, temveč na večje enote, ki vključujejo več fonetičnih segmentov, kot so besede, fraze, stavki ali celo večji odseki govorjenega besedila. Zato o prozodičnih lastnostih govora pogosto govorimo kot o nadsegmentnih prozodičnih lastnostih (Gros 1997). Tekoči govor strukturirajo v prozodične fraze, jakostne izrazitosti v povedi in intonacijsko glavo. Zaznavamo jih kot: naglas, spremembo v intonaciji, ritem in glasnost.

Domena izboljšanja napovedovanja simbolične prozodije ostaja eden večjih izzivov sistemov TTS. Pristopi modeliranja prozodije, ki se uporabljajo v večini sistemov za TTS, so ponavadi preprosti. Sistemi TTS, ki podpirajo napredno modeliranje intonacije, ritma in poudarjenosti besed kar se da verodostojno, so redki. Še bolj redki so sistemi, ki vključujejo ustrezne modulacije osnovne harmonske frekvence (f_0) na besedni ravni (Gros, 1997).

Visokokakovostno sintezo govora pogojuje ustrezno modeliranje prozodije za določanje strukture prozodičnih fraz, identifikacijo jakostnih izrazitosti v frazah in določanje trajanja posameznih fonemov (Vereecken et al., 1998). Hkrati moramo izpostaviti dejstvo, da je modeliranje prozodije zelo kompleksen proces. Dejstvo, da se je bolje izogniti modeliranju prozodije, kot pa jo modelirati slabo (Terken, 1995), nas je vodilo h konceptu večnivojskega (hierarhičnega) pristopa modeliranja prozodije – ločeno na segmentnem in nadsegmentnem nivoju. V nadaljevanju bomo predstavili del postopka modeliranja na nadsegmentnem nivoju, ki se nanaša na napovedovanje prozodičnih mej.

2. Modeliranje simbolične prozodije

Pri tvorjenju simbolične prozodije sta običajna dva pristopa (Fackrell, 1999);

- pristop oblikovanja jezikovnega modela z lingvističnim ekspertnim znanjem in ročnim modeliranjem, ki pogojuje tako ekspertno znanje kot togost modela, vezanega na specifični jezik in ciljno procesiranje, ter
- pristop s samodejnimi tehnikami učenja z ustreznimi podatkovnimi bazami, ki predstavljajo temelj podatkovno vodenim pristopom. Slednje odlikujeta potencial hitrega razvoja modelov in jezikovna neodvisnost – podpora večjezičnosti.

Postopki podatkovno vodenih pristopov modeliranja simbolične prozodije, ki predstavljajo enega ključnih gradnikov sodobnih sistemov TTS, pogojujejo ustrezno besedilno zbirko. V splošnem govorimo v kontekstu simbolične prozodije o označevanju prozodičnih mej

(členitev besedila na fraze) in označevanju jakostnih izrazitosti v povedi (Toporišič, 1996).

Postopek podatkovno vodenega pristopa modeliranja simbolične prozodije lahko razčlenimo na:

- korak določanja abstraktnega simboličnega opisa na osnovi akustične strukture povedi oz. manjših segmentov (samodejno/ročno), lahko tudi z upoštevanjem skladenjske strukture povedi (npr. Flach, 1999, Mihelič, 1999),
- korak napovedovanja simboličnih značnic, izločenih iz samodejno/ročno označene besedilne zbirke z eno od tehnik samodejnega učenja (npr. Malfrere, 1998), in
- korak preslikave simboličnega niza značnic v akustično interpretacijo, npr. s superponiranjem dobljene prozodične hierarhije mikroprozodičnemu jedru (modeliranje trajanja fonemov, poteka osnovne harmonske frekvence posameznih segmentov ter globalne intonacijske krivulje povedi z upoštevanjem jakostne izrazitosti besed in trajanja premorov med frazami) (Stergar, 2000).

3. Podatkovno vodeni pristop

Podatkovno vodene tehnike pogojujejo zasnovo obsežne besedilne zbirke, ki jo moramo večinoma označiti ročno. Proces označevanja besedilne zbirke je izjemno obsežen in zahteva specifično ekspertno znanje označevalcev, hkrati pa pogojuje nekonsistentnost pri označevanju (velikokrat je za označevanje obsežne besedilne zbirke potrebnih več ekspertov). Nenazadnje je proces ročnega označevanja skrajno zamuden in drag, kar predstavlja oviro pri hitri adaptaciji sistema TTS na prozodične značilnosti govorca oz. specifičen jezik.

Zaradi v predhodnem odstavku omenjene problematike je v literaturi vse več sugestij pristopov označevanja besedilnih zbirk s samodejnimi postopki (Vereecken et al., 1998, Malfrere et al., 1998). Kljub temu pa avtorji opozarjajo, da je potrebno zaradi doslednosti pri označevanju pregledati označeni material in ga po potrebi dopolniti. Z ustreznimi postopki pri označevanju (npr. uporabo specifičnega orodja za označevanje) se konsistentnost označevanja poveča, časovni obseg označevanja pa bistveno zmanjša (Stergar, 2000).

3.1. Uporabljena besedilna zbirka

Uporabili smo besedilno zbirko, zasnovano za konkatenativno sintezo slovenskega jezika (Rojc, 2000, Stergar, 2000). Besedilna zbirka je posneta v studijskem okolju in obsega pribl. 3 ure branega slovenskega govora (pribl. 1200 povedi). Povedi so normirane in vsebujejo fonemski prepis (transkripcijo glasov). Vse so bile izbrane iz obsežne besedilne zbirke (pribl 2 milijona povedi) zajete iz literature in dnevnega časopisja v elektronski obliki, večinoma v povednem naklonu. Povedi, ki obsegajo med 15 in 25 besed je izolirano interpretiral profesionalni radijski napovedovalec. Večina povedi je v povednem naklonu z dinamično interpretacijo prozodije.

Oblikoslovno analizo povedi v besedilni zbirki smo izvedli ročno (POS). Uporabili smo naslednje oznake:

1. SUBST za samostalniške besede z izjemo samostalniških zaimkov
2. VERB za glagole
3. ADJ za pridevnike
4. ADV za prislove
5. NUM za vse vrste številnikov
6. PRON za zaimke (samostalniške in pridevniške)
7. PRED za povedkovnik
8. PREP za predloge
9. CONJ za veznike
10. PART za členke
11. INT za medmete
12. EPUNC za končna ločila ter : in ;
13. IPUNC za vsa ostala ločila

3.2. Označevanje prozodičnih mej

V splošnem se uporabljata za označevanje prozodičnih mej dva nabora značnic (Stergar, 2000):

- nabor za označevanje skladenjskih prozodičnih mej (sintaktično-prozodičnih mej) in
- nabor za označevanje akustičnih prozodičnih mej (prozodičnih mej).

Raziskave kažejo, da so sintaktično-prozodične meje podmnožica prozodičnih in ponavadi z njimi ne sovpadajo oz. pokrivajo le del prozodičnih mej (Terken, 1995, Flach, 1999). Nanje vplivata razen skladnje tudi ritem in specifičen slog interpretacije govorca. Za brano angleščino pokrivajo sintaktično-prozodične meje zgolj 65 % prozodičnih mej (Flach, 1999).

V prvi fazi označevanja prozodičnih mej v besedilni zbirki smo iz v predhodnem odstavku navedenih razlogov pristopili k označevanju na osnovi akustičnih pokazateljev.

3.2.1. Nabor značnic za označevanje prozodičnih mej

Inventar značnic za označevanje prozodičnih mej – prozodičnih značnic – smo zasnovali na obstoječih inventarjih (Kompe, 1997, SI1000, 1998, Mihelič, 1999, Batliner, 1997). Definirali smo naslednje značnice:

- B3 omejuje zaključene, poudarjene fraze. Kot poudarjeno frazo smo označili skupino besed, od katerih nosi vsaj ena primarni naglas (normalno naglašena beseda). Značnica B3 mora biti po definiciji dvotonska (bitonalna), saj jo morata sestavljati frazni poudarek in mejni ton. Velikokrat označuje tudi mejo, določeno s premorom.
- B2 nastopa znotraj fraz, označenih z B3, in označuje podrobnejšo prozodično podstrukturo – podfrazo. Podfraza je intonacijsko šibkejša kot njej nadležeča prozodična struktura.
- B9 označuje nepravilno členjene meje, ki nastanejo zaradi nenamernih premorov pri izgovarjavi, oklevanju ipd. in v bistvu nimajo bistvene funkcije v prozodični strukturi (SI1000, 1998). Z B9 smo zaenkrat označili vse prozodične meje, ki niso označene z B3, B2 ali B0.
- B0 za vse ostale meje.

3.2.2. Postopek določanja prozodičnih mej

Za določanje prozodičnih mej smo zasnovali posebno grafično orodje (Stergar, 2000). Besedilno zbirko smo označili v dveh korakih. V prvem smo samodejno določili vse premore, pri čemer smo uporabili orodje za segmentiranje HTK. Premore smo razvrščali glede na mesto nastopanja; znotraj povedi in med povedmi (končna ločila). Dolžino trajanja posameznih premorov glede na

stavčno strukturo zaenkrat nismo analizirali. V drugem koraku smo ročno popravili razvrščene prozodične meje in besedilno zbirko dopolnili s simboličnimi oznakami za prozodične meje, ki niso pogojene s premori (baza PB2). Pri tem smo med akustičnimi pokazatelji izpostavili potek osnovne harmonske frekvence. V obeh korakih smo označili polovico besedilne zbirke (600 povedi) (Stergar, 2000).

4. Napovedovanje prozodičnih mej

Za napovedovanje simbolične prozodije so razširjeni različni pristopi napovedovanja, npr. nevronske mreže, prikriti modeli Markova, binarna odločitvena drevesa (Black, 1997, Fackrell, 1999). V novejših poskusih napovedovanja prozodičnih značnic (Müller, 2000) pa je zaznaven trend uporabe kompleksnejših struktur nevronskih mrež.

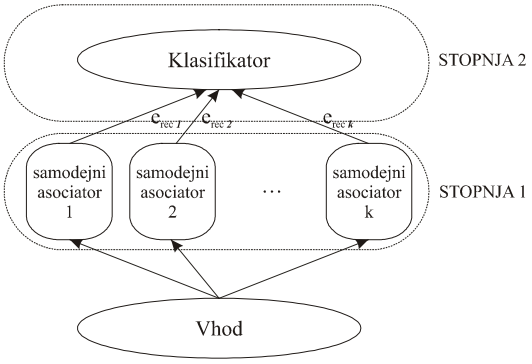
4.1. Uporabljena arhitektura nevronskih mrež

Težava, s katero se srečujemo pri uporabi večplastnih (ena prikrita plast) perceptronov (MLP) z velikimi dimenzijami vhodnih vektorjev, je neuravnotežen pretok informacije med povratnimi in vhodnimi podatki. Informacija v vhodnem vektorju se za klasifikacijo napake na izhodu preslika preko prikrite plasti v zgolj eno samo dimenzijo (spremembe v vhodnem vektorju velikih dimenzij se zrcalijo zgolj v eni sami vrednosti).

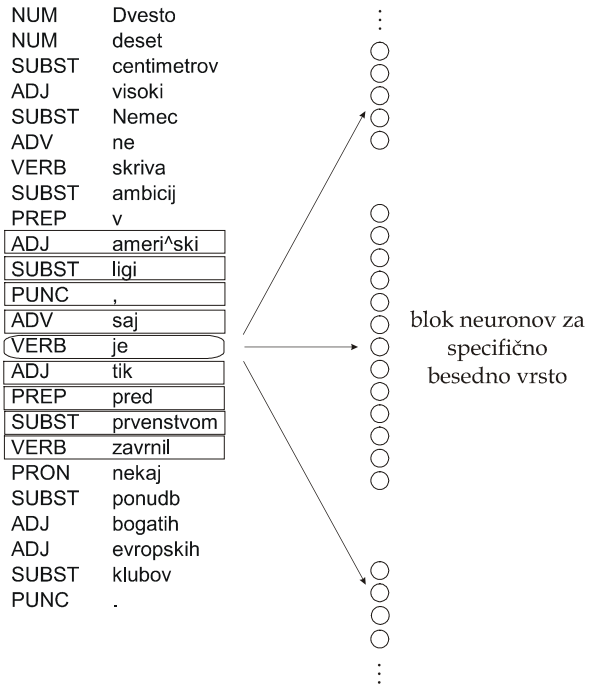
Z uporabljeno arhitekturo smo se skušali izogniti omenjenemu problemu tako, da smo proces učenja razčlenili na dva koraka – korak ločenega učenja modelov za posamezen razred (vsaka značnica za označevanje prozodičnih mej predstavlja po en razred) in korak klasifikacije verjetnosti iz rekonstrukcijske napake posameznih modelov (Slika 1).

4.2. Vhodni podatki

Kateri parametri so relevantni za napovedovanje simboličnih značnic, ostaja odprto vprašanje. Ustrezen izbor značilk lahko bistveno pripomore k uspešnosti napovedovanja (Hirsberg, 1993, Ostendorf, 1994), vendar je le-ta pogojen z ustreznim lingvističnim ekspertnim znanjem. Na tak način dobljena množica značilk pa je lahko odvisna tako od jezika in tudi od ciljne aplikacije. Značilke, ki se razširjeno uporabljajo in so navidez neodvisne od jezika in ciljne aplikacije, so oblikoslovna zaporedja (nizi besednih vrst), dobljena z oblikoslovno analizo povedi (POS) (Müller, 2000).



Slika 1: Dvostopenjska arhitektura za klasifikacijo napake samodejnih asociatorjev.



Slika 2: Preslikava vhodne informacije v vhodno plast AAC.

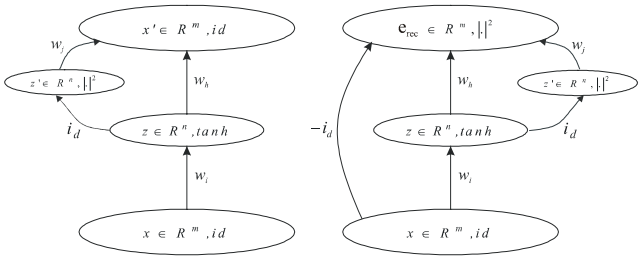
Vhodne vzorce smo razvrstili v simetrično strukturo s širino niza 9 (Slika 2). Za obravnavano besedo (za katero želimo napovedati prozodično mejo in se nahaja v sredini niza) smo upoštevali levo in desno zaporedje besednih vrst dolžine 4. Dolžina vhodnega vektorja je določena s širino niza in obsegom deklariranih besednih vrst. V našem primeru je vhodni vektor $m = (4 + 1 + 4) * 13 = 117$ (m je število nevronov v vhodni plasti).

Vrednosti za posamezno besedno vrsto smo zakodirali ternarno (1 = aktiven vhod, 0 = neveljaven vhod in –1 = neaktiven vhod).

4.3. Samodejno-asociativna nevronska mreža

Arhitektura za klasifikacijo napake samodejnih asociatorjev pogojuje ločeno učenje za vsak razred. m-dimenzionalni vhodni vektor preslikamo v n-dimenzionalni vektor z, kjer je $n << m$. Nevronsko mrežo učimo s ciljem čim boljše adaptacije izhodnega vektorja x' na vhodni vektor x. Podatki so v vmesni plasti nevronske mreže najprej predstavljeni zgoščeno preko matrične preslikave $w_1 \in \Re^{nxm}$ in nato razredčeno v izhodni plasti preko matrike $w_2 \in \Re^{mxn}$ (Slika 3, levo). Učinkovitost strukture lahko izboljšamo z dodatno plastjo (z^2), ki povezuje srednjo, zgoščevalno plast z izhodno (Neuneier, 1998).

Po prvem koraku učenja uporabimo za vsak posamezen razred (stopnja 1) strukturo nevronskih mrež za ugotavljanje rekonstrukcijske napake e_{rec} (Slika 3, desno). Negativna enotina matrika $-id \in \Re^{mxm}$ in kvadratična aktivacijska funkcija $|.|^2$ v izhodni plasti predstavljata vektor razdalje $e_{rec} = (x' - x)^2$, ki podaja podrobno informacijo o napaki, potrebni za ugotavljanje pripadnosti določenemu razredu prozodičnih značnic v koraku klasifikacije (npr. B3/B2). Eksperimenti potrjujejo, da lahko z omenjeno arhitekturo izločimo značilnosti različnih razredov.



Slika 3: Levo: samodejni asociator za ločeno učenje razredov (stopnja 1). Desno: Samodejni asociator za določanje vektorja napake po prvi stopnji učenja.

5. Rezultati

Izvedli smo poskus uspešnosti napovedovanja prozodičnih mej z uporabo ročno dopolnjene in popravljene baze PB2 (razdelek 3.2.2). Prozodične meje, označene z B3, smo definirali kot pomembnejše, prozodične meje B2 pa kot manj pomembne za označevanje intonacijsko šibkejše prozodične podstrukture v povedi (razdelek 3.2.1). V poskusih napovedovanja prozodičnih mej, navedenih v nadaljevanju (Tabela 1), zaenkrat nismo obravnavali prozodičnih mej ločeno (B2 = B3).

Tabela 1: Uspešnost napovedovanja za B2 = B3.

prozodične meje	PB	NB	splošno
B2 = B3	89,88 %	4,74 %	94,55 %

Tabela 2: Matrika zamenjav prozodičnih mej.

značnica/napovedana	PB	NB	vse napovedane meje
PB	506	174	680
NB	57	3498	3555
vse meje v bazi	563	3672	

Z oznako PB smo označili vse pravilno napovedane prozodične meje, z NB smo označili vse nepravilno napovedane (na mestih, kjer ni meje, je bila meja napovedana) (Tabela 1). Za kriterij uspešnosti napovedovanja prozodičnih mej smo uporabili odstotek pravilno napovedanih mej, odstotek splošne pravilnosti napovedovanja (za PB in NB) in odstotek nepravilno napovedanih NB (Black, 1997). Splošna pravilnost napovedovanja znaša 94,6 %, v najslabšem primeru, če ne bi uspeli napovedati nobene prozodične meje pravilno, pa bi ta rezultat znašal 82,6 %. Omenjen podatek smo navedli zaradi primerjave splošne uspešnosti napovedovanja prozodičnih mej. Število prozodičnih mej v primerjavi s številom mej kjer prozodične meje ne nastopajo je zelo neuravnoteženo (PB : NB ≈ 1: 9). Če bi (zgolj hipotetično) uspešno napovedali vsa mesta kjer prozodične meje ne nastopajo (in nobene prozodične meje) bi znašala splošna uspešnost napovedovanja kot navedeno.

Če primerjamo uspešnost napovedovanja prozodičnih mej (89,9 %, Tabela 1) s predhodnimi poskusi napovedovanja (Stergar, 2000) lahko izpostavimo bistven napredek v primerjavi s klasičnim napovedovanjem z MLP pri nespremenjeni neuspešnosti napovedovanja NB (mest kjer prozodične meje ne nastopajo). Izboljšanje uspešnosti napovedovanja je rezultat nove strukture nevronskih mrež, ki ni tako dovzetna na asimetrijo vhodnih podatkov pri učenju (PB : NB) in tako tudi bistveno pripomore k splošni uspešnosti napovedovanja prozodičnih mej.

6. Zaključek

V članku smo predstavili prve poskuse napovedovanja prozodičnih mej z arhitekturo klasifikacije napak samodejno-asociativnih nevronskih mrež. Arhitekturo smo preskusili s ciljem ustreznosti za napovedovanje simboličnih prozodičnih mej v kontekstu modeliranja prozodije sistemov TTS. Rezultati uspešnosti razvrščanja prozodičnih mej z uporabo omenjene arhitekture so obetavni in kažejo na primernost strukture za napovedovanje prozodičnih mej. Predpostavljamo, da je k obetavnemu rezultatu uspešnosti napovedovanja prispevala tudi uporaba grafičnega orodja in samodejnega postopka označevanja, ki sta bistveno prispevala h konsistentnosti pri označevanju, hkrati pa se zavedamo, da so za potrditev naše hipoteze potrebne nadaljnje raziskave.

7. Reference

Black A. W., Taylor P. (1997). Assigning Phrase Breaks from part-of-speech sequences. Eurospeech 97'. Rodos, Greece, 1997.

Edgington M., Lowry A., Jakson P., Breen A. P., Minnis S. (1996). Overview of current Text.-to-Speech techniques. BT Technology Journal Vol 14 No 1 January 1996.

Fackrell J. W. A., Vereecken H., Martens J.-P., Van Coile B. (1999). *Multilingual Prosody Modelling using Cascades of Regression trees and Neuronal Networks*. Eurospeech'99. Budapest, Hungary. 1999.

Flach M. L. (1999). *A comparison between syntactic and prosodic phrasing*. Eurospeech '99, Budapest, Hungary. 1999.

Gros J. (1997). Samodejno tvorjenje govora iz besedil.. Doktorska disertacija. Univerza v Ljubljani. Fakulteta za elektrotehniko. 1997.

Hayes B. (1995). Metrical Stress Theory: Principles and Case Studies. Chicago: University of Chicago Press. 1995.

Hirsberg J. (1993). Pitch accent in context: Predicting prominence from text. Artificial Intelligence, 1993.

Kompe R., (1997). *Prosody in Speech Understanding Systems*. Springer – Verlag Berlin Heidelberg, Lecture Notes in Artificial Inteligence 1307, 1997.

Malfrere F., Dutoit T. and Mertens P. (1998*). Fully automatic prosody generator for text-to-speech*. ICSLP 98, Sydney Australia. 1998.

Mihelič F., Gros J. (1999). Recognition of Prosodic Events in Slovenian Speech. ERK 1999 Portorož, Slovenia. 1999.

Müller A. F., Zimmermann H. G., Neuneier R. (2000). Robust generation of symbolic prosody by a neural classifier based on autoassociators. IEEE ICASSP 00'. Istanbul, Turkey, 2000.

Neuneier R., Zimmermann H. G. (1998). How to train neural networks. In G. B. Ohrr & K.-R. Müller, editors, Neural Networks: Tricks of the trade. Springer Verlag, Berlin. 1998.

Ostendorf M., Veilleux N. (1994). A hierahical stohastic model for automatic prediction of prosodic boundary location. Computational Linguistics. 1994.

Rojc M., Kačič Z. (2000). *Design of Optimal Slovenian Speech Corpus for use in the concatenative Speech Synthesis System.* LREC'00 Athens Greece. 2000.

Selkirk E. O. (1984). Phonology and Syntax. Cambridge, MA: MIT Press. 1984.

SI1000 (1998) – Prosodic Markers Version 1.0, Bavarian Archive of Speech Signals. University of Munich, Institute of Phonetics, Germany. 1998.

Stergar J. (2000). Določanje prozodičnih značilnosti z analizo besedil. Magistrsko delo. Univerza v Mariboru. Fakulteta za elektrotehniko, računalništvo in informatiko. 2000.

Terken J., Collier R. (1998). The Generation of Prosodic Structure and Intonation in Speech Synthesis in W. B. Klein et all eds. Speech Coding and Synthesis. Elsevier. 1998.

Toporišič J. (1996). Slovenska slovnica. Založba Obzoja Maribor. 1996.

Vereecken H., Martens J. P., Grover C., Fackrell J., Van Coile B. (1998). *Automatic prosodic labeling of 6 languages*. ICSLP 98, Sydney Australia. 1998.