ELEKTROTEHNIŠKI VESTNIK 80(4): 171-176, 2013
IZVIRNI ZNANSTVENI ČLANEK

Enostavnejša zasnova sistema za razpoznavanje govora

Robert Rozman

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Tržaška 25, 1000 Ljubljana, Slovenija
E-pošta: robert.rozman@fri.uni-lj.si

Povzetek. V delu so predstavljene glavne slabosti obstoječih sistemov za razpoznavanje govora. Ti so namreč dokaj statični sistemi z veliko vnaprej vgrajenega znanja, ki se med delovanjem (pre)malo spreminja. Zato je treba povečati količino iz podatkov pridobljenega znanja, kar je še posebej pomembno za manjše jezikovne prostore. Pri tem so še zlasti pomembna skupina kompaktni sistemi za razpoznavanje govora z enostavnejšo zasnovo in nevronsko mrežo v vlogi akustičnega modela. Po našem mnenju je zasnova teh sistemov bolj kompatibilna z najnovejšimi dosežki na področjih porazdeljenega in vzporednega procesiranja. Predstavljena sta dva praktična sistema: prvi je besedni razpoznavalnik z omejenim slovarjem in nevronsko mrežo kot akustičnim in postopkom Viterbijevega iskanja kot časovnim modelom; drugi pa je enostavnejši sistem z enakim akustičnim, a brez časovnega modela. Njun osnovni namen je podpora nadaljnjim raziskavam na tem področju.

Ključne besede: govor, fonem, razpoznavanje govora, jezikovne tehnologije

Simplified design of the Speech Recognition System

Disadvantages of the currently used Speech Recognition Systems (SRSs) and alternative ways of their evolution are presented and discussed. In our opinion, SRSs are rather static structures with a lot of predefined knowledge that is built into them upon their creation, and usually remaining unchanged or unadapted during the recognition process. Several possible ways of increasing the amount of the dynamic, automatically learned knowledge in the next generation of SRSs are discussed; this is of a particular importance for under-resourced languages. A group of SRSs, i.e. compact SRSs with a limited vocabulary based on the Neural Network as an acoustic model, is of a particular interest. Its structure is more compatible with the recent developments in the field of distributed and parallel processing.

Two experimental systems are presented and tested on a simple phoneme recognition task. One system is a fairly complete SRS based on the Neural Network as an acoustic model and Viterbi search as a time model. The second system is much simpler using only the Neural Network as an acoustic model. It will support our further research in this field.

1 UVOD

Sistemi za razpoznavanje govora (v nadaljevanju SRG) se že kar nekaj časa intenzivno razvijajo; od začetnih, najenostavnejših sistemov z nekaj besedami v slovarju, pa vse do današnjih, sodobnejših strežniških sistemov s kompleksnimi slovarji in gramatikami oziroma jezikovnimi modeli, ki omogočajo razpoznavanje tekočega govora poljubnega govorca.

Kljub intenzivnemu razvoju, kontinuirani širitvi slovarjev in uporabi čedalje bolj zapletenih jezikovnih modelov na konceptualni ravni razpoznavanja govora že nekaj časa ni zaznati bistvenih sprememb. V zadnjem času je razvoj čedalje bolj pod vplivom dosežkov s področja aparaturne opreme – predvsem naraščanja zmogljivosti s pomočjo vzporednega (večjedrni procesorji) oziroma porazdeljenega procesiranja (zmogljive mrežne povezave). Tako so današnji SRG zmogljivejši, sestavljeni iz več vzporednih (tudi oddaljenih) enot, ki jih lahko izkoristimo za učinkovitejše razpoznavanje predvsem na dva načina:

- opravljamo več hkratnih, konceptualno različnih razpoznav istega govornega signala, katerih izide zlijemo v enoten rezultat;
- opravljamo več hkratnih, preprostejših in med seboj popolnoma neodvisnih razpoznav različnih govornih signalov.

V drugem primeru lahko prednosti vzporednega in porazdeljenega procesiranja precej bolj učinkovito izkoristimo. V obeh primerih pa lahko enakomerneje porazdelimo tudi računsko breme – uporabniku najbližje mobilne naprave zaradi omejene zmogljivosti prevzamejo le sorazmerno manjši del celotne obdelave oziroma razpoznave govornega signala, preostali del pa se dodeli bolj zmogljivim, oddaljenim strežnikom.

Dosedanji razvoj na področju razpoznavanja govora ima tudi nekatere negativne posledice. V obstoječih SRG je vgrajenega precej statičnega znanja, ki se ob delovanju sistema le malo spreminja in prilagaja trenutnim razmeram; tipičen primer je proces parametrizacije.

Iz povedanega sledi, da se bosta v prihodnosti še naprej spreminjala tako zasnova kot tudi način delovanja SRG. Zaradi razvoja na drugih, sorodnih področjih, lahko rečemo, da se tudi pri razpoznavanju govora nahajamo na pomembni razvojni prelomnici.

Prejet 30. september, 2013
Odobren 11. november, 2013

V nadaljevanju najprej predstavljamo nekaj po našem mnenju osnovnih slabosti obstoječih SRG, ki bi jih sodobnejše zasnove morale odpraviti. Nato opišemo dva eksperimentalna sistema, ki sta korak v smeri enostavnejše zasnove SRG. Predstavljeni so tudi rezultati preliminarnih preizkusov na obeh sistemih.

2 ALTERNATIVNE SMERI RAZVOJA SISTEMOV ZA RAZPOZNAVANJE GOVORA

Pri lastnih raziskavah se že nekaj časa osredinjamo na alternativne možnosti nadaljnjega razvoja SRG. Med njimi je po našem mnenju zelo pomembno področje kompaktnih SRG z omejenim slovarjem, ki temeljijo na enostavnejši zasnovi in razpoznavanju bolj elementarnih govornih enot – npr. fonemov, fonemskih podkategorij.

Za realizacijo tovrstnih sistemov uporabljamo orodje CSLU Speech Toolkit [1], ki omogoča realizacijo razpoznavalnikov na podlagi preproste večnivojske nevronske mreže (MLP, »Multi Layer Perceptron«) kot akustičnega modela; ta koncept je računsko precej manj zahteven kot kompleksnejši SRG. Ti so sicer namenjeni reševanju težjega problema – razpoznavanju tekočega govora poljubnega govorca z velikim številom besed v slovarju, vendar imajo prav zaradi zahtevnosti problema tudi nekatere slabosti:

- večja računska kompleksnost,
- slabša skalabilnost (ob večjem številu hkratnih razpoznav),
- slabše prilagajanje individualnim značilnostim govorca,
- potreba po obsežnih in specifičnih govornih zbirkah, potrebnih za učenje.

2.1 Manj zapletenih ali več enostavnih SRG?

Glede na naše izkušnje menimo, da lahko prej omenjeni enostavnejši (kompaktni) SRG z omejenim slovarjem na tem področju še vedno obdržijo pomembno vlogo in bodo vsaj na konceptualni ravni prisotni tudi v prihodnje. Njihove glavne prednosti so ravno pravi odgovor na prej opisane slabosti kompleksnejših SRG:

- manjša računska zahtevnost; izvajajo se lahko tudi na mobilnih napravah,
- lažja integracija v vzporedne in porazdeljene sisteme procesiranja,
- zaradi neodvisnosti posameznih razpoznav je skalabilnost boljša,
- učinkovitejša prilagoditev končnemu uporabniku (tudi individualnim značilnostim).

Naštete prednosti v zadnjem času postajajo še bolj izrazite zaradi razvoja na področju IKT in s tem povezanih primerov množične uporabe:

- pametnih telefonov in z njimi razpoznavanja govora v vsakdanjem življenju; v tem primeru so slovarji bolj omejeni,
- porazdeljenih in vzporednih procesorskih sistemov, ki omogočajo učinkovitejšo realizacijo večjega števila manjših, enostavnejših SRG.

V zadnjem času se intenzivno razvija še nekaj za razpoznavanje govora zelo pomembnih področij; med njimi so posebno zanimiva področja razpoznavanja vzorcev, strojnega učenja in podatkovnega rudarjenja. Z njimi se čedalje bolj prepleta tudi razpoznavanje govora. Z nekaj poenostavitvami ga lahko namreč obravnavamo kot splošnejši problem, rešljiv tudi z metodami s teh sorodnih področij. Zato lahko na vseh omenjenih področjih pričakujemo sinergične učinke in hitrejši napredek. Dolga leta je bilo namreč razpoznavanje govora precej osamljeno in so bile dosežene rešitve preveč statične in prilagojene posebnostim na tem področju; s tem se je nehote omejeval tudi nadaljnji razvoj. Sodoben multidisciplinarni razvoj pod vplivom sorodnih področij lahko ponudi splošnejše rešitve, ki so potencialno zanimivejše; statično znanje se bo tako laže nadomestilo s tistim, pridobljenim po avtomatski poti, v največji meri iz obstoječih podatkov. Ker so splošnejše metode reševanja uporabne na širšem spektru sorodnih problemov, lahko ob nižjih stroških razvoja in delovanja sistemov pridemo do boljših končnih rezultatov. To dejstvo je še zlasti pomembno za jezike, ki so glede razpoložljivih virov za njihov razvoj omejeni – mednje vsekakor spada tudi slovenski jezik. Naj poudarimo čedalje bolj sprejeto dejstvo, da bo obstoj jezikov v precejšnji meri odvisen od ustrezne podpore v sodobnih jezikovnih tehnologijah.

2.2 Sodobna dinamična zasnova SRG

V opisani smeri potekajo tudi naše raziskave, ki poskušajo problem razpoznavanja govora predstaviti v nekoliko bolj splošni in za začetek tudi enostavnejši obliki, ki bo bolj uporabna tudi z vidika že omenjenih sorodnih področij avtomatskega pridobivanja znanja iz podatkov. Posledica teh prizadevanj je lahko fleksibilnejša zasnova SRG kot resna alternativa ustaljeni zgradbi tovrstnih sistemov, ki je prevladovala doslej. Tipičen kazalec nefleksibilnosti zasnove večine obstoječih SRG so različne oblike vnaprej vgrajenega znanja, ki so se med razvojem le malo spreminjale:

- statična zgradba SRG, ki se med delovanjem bolj malo spreminja,
- standarden postopek tvorbe opisa govornega signala – t. i. parametrizacija,
- »klasičen« nabor metod učenja iz govornih zbirk oziroma podatkov, ki večinoma vsebujejo določene omejitve (npr. nediskriminatorsko učenje) ali pogosto dvomljive predpostavke o lastnostih podatkov (statistična porazdelitev, neodvisnost vektorjev značilk)

Kot alternativa opisani statični zgradbi SRG se pojavlja bolj odprta zasnova razpoznavalnika preprostejših govornih enot (npr. fonemov, fonemskih podkategorij) v kombinaciji z novejšimi koncepti oziroma metodami z že omenjenih področij avtomatskega učenja. Na ravni preprostejših govornih enot je sicer razpoznava manj uspešna, vendar je zelo hitra in učinkovita. Tipično število mogočih hipotez je

namreč bistveno manjše od števila besed v slovarjih kompleksnejših SRG.

S prehodom na preprostejše govorne enote in bolj »čisto« zasnovo se sistem laže povezuje z drugimi sorodnimi sistemi in višjimi ravnmi obdelave informacij, ki niso več izključno del istega sistema. Tako se manjša uspešnost na ravni preprostejših govornih enot lahko nadomesti na višji ravni besednega in jezikovnega modeliranja – med drugim tudi zaradi lažje uporabe splošnejših metod iz že omenjenih sorodnih področij.

Morda najpomembnejši vidik tovrstne spremembe v pristopu je ta, da postane problem razpoznavanja preprostejših govornih enot precej bolj podoben drugim problemom s sorodnih področij (slike, vzorci) in je potemtakem rešljiv s splošnejšimi metodami. S tem lahko zmanjšamo količino vnaprej določenega, statičnega znanja in ga nadomestimo z avtomatsko naučenim znanjem iz podatkov, ki je lahko precej bolj dinamično in bolje prilagojeno trenutnim podatkom ter manj odvisno od konkretnega problema.

Na področju razpoznavanja govora je v zadnjem času mogoče zaslediti več tovrstnih alternativnih pristopov. Med njimi so za naša izhodišča zanimivi zgodnji poskusi fonemske oziroma segmentne zasnove razpoznavanja; te še najbolj konsistentno sledijo že opisani alternativni paradigmi. Med njimi sta po našem vedenju pomembna predstavnika sistema SUMMIT iz MIT [2] in ROAR iz CSLU [3], pojavljajo pa se tudi novejše različice omenjenih pristopov (npr. [4]). V zadnjem času so bili doseženi zelo obetavni rezultati na tem področju tudi s t. i. globokimi mrežami (DBN-»Deep Belief Networks«) [5]. Koncept se intenzivno razvija, zato je za celovito oceno še prezgodaj. Vsekakor pa so glede na rezultate globoke nevronske mreže pravi korak razvoja v smeri avtomatskega učenja in dinamičnega prilagajanja obstoječim podatkom.

Z vsemi naštetimi alternativnimi pristopi pa se je zgodilo še nekaj zanimivega – nižje ravni razpoznavanja so spet pridobile na pomembnosti. Dolga leta se namreč s procesi, kot sta parametrizacija in frekvenčna analiza, večina raziskovalcev razen nekaj izjem [6] na tem področju ni intenzivno ukvarjala.

V nadaljevanju so najprej predstavljene metode in nato še preliminarni preizkusi, opravljeni z namenom evalvacije perspektive tovrstnega alternativnega pristopa. Z izbiro govorne zbirke smo dali ustrezen poudarek slovenskemu jezikovnemu prostoru. Vsaj enako pomemben pa je tudi poudarek na vzpostavitvi eksperimentalnega okolja za nadaljnje raziskovalno delo na tem področju.

3 FONEMSKO RAZPOZNAVANJE GOVORA

V skladu z opisanimi izhodišči smo opravili preliminarne praktične preizkuse na dveh različnih sistemih, ki temeljita na razpoznavanju preprostejših govornih enot – fonemov.

Prvi sistem je pravzaprav celovit SRG z omejenim slovarjem besed (v nadaljevanju CSLU-SRG), ki temelji na uporabi nevronske mreže v vlogi akustičnega modela; že nekaj časa ga uporabljamo pri vseh naših raziskavah.

Drugi sistem je preprostejši razpoznavalnik (klasifikator) fonemov, ki za vsak okvir določi verjetnost pripadnosti razredom vseh fonemov. Sistem je zasnovan kot razvojno oziroma eksperimentalno okolje, v katerem lahko preizkušamo različne ideje oziroma alternativne pristope in ga ob tem učinkovito dopolnjujemo.

V nadaljevanju opisujemo najprej tri skupna implementacijska izhodišča in nato še ločeno podrobnosti realizacije obeh sistemov.

3.1 Parametrizacija – postopek tvorbe opisov govornega signala

Postopek parametrizacije v obeh testnih SRG je bolj ali manj enak splošno znanemu postopku računanja t. i. značilk MFCC [7].

Govorni signal se razdeli na 32 ms dolge okvirje z 10 ms medsebojnega razmika. Za vsak okvir se izračuna frekvenčna predstavitev signala. Ker je ta preveč podrobna, se združi v bistveno manjše število frekvenčnih oziroma t. i. kritičnih pasov (22), ki so razporejeni po nelinearni melodični lestvici (angl. Mel-Scale) v skladu z lastnostmi človekovega sluha. Ker se kritični pasovi med seboj prekrivajo, so značilke redundantne in med seboj korelirane. Zato se običajno izvede še izračun značilk MFCC, ki koreliranost in redundanco zmanjšajo. V našem primeru smo izračunali 13 značilk MFCC, ki sestavljajo osnovni vektor oziroma opis signala v posameznem okvirju. Po navadi se v praksi osnovnemu vektorju dodajo še značilke delta, ki opisujejo spremembe vrednosti osnovnih značilk v nekaj sosednjih okvirjih. Namesto teh smo pri obeh sistemih uporabili t. i. kontekstno okno; več sosednjih okvirjev smo preprosto združili v skupni vektor 65 značilk (poleg vektorja v času t so prisotni še vektorji pri t-60ms, t-30ms, t+30ms, t+60ms). S tem pri klasifikaciji upoštevamo dinamiko daljšega časovnega intervala (v tem primeru 120 ms) v primerjavi s klasičnimi značilkami delta, ki po navadi opisujejo le zadnjih 50 ms signala. Na končnih značilkah se je izvedel še standardni postopek normalizacije – odštevanja povprečne vrednosti na ravni posameznih izgovarjav (CMN–»Cepstral Mean Normalization«). Več podrobnosti o postopku parametrizacije bralec najde v [7] in drugi literaturi s področja razpoznavanja govora.

Kot smo že omenili, je parametrizacija tipičen primer nefleksibilnosti obstoječih SRG in vnaprej vgrajenega znanja, ki se med delovanjem sistema ne spreminja. Nekatere naše pretekle raziskave so pokazale ([6],[8],[9]), da bi se v tej smeri splačalo intenzivneje raziskovati; to dejstvo pa potrjujejo tudi najnovejše raziskave, ki so po daljšem času spet bolj intenzivne na nižjih ravneh razpoznavanja [5]. Zato bo proces

parametrizacije eno glavnih področij naših nadaljnjih raziskav.

3.2 Kontekstna odvisnost fonemov

Nevronske mreže v obliki MLP so za modeliranje časovnih zaporedij dokaj neprimerne. Zato realizacija besednih modelov z njimi ni smotrna. Logična izbira so po akustičnih in časovnih značilnostih bolj homogene enote – fonemi. Zato se besede v slovarju zapišejo kot zaporedja fonemov, ki ponazarjajo časovni potek njihove izgovarjave.

Izgovarjava fonema je po navadi odvisna od predhodnega in naslednjega fonema. Pojav se imenuje koartikulacija in je za govor zelo značilen; treba ga je upoštevati tudi pri modeliranju fonemov. Običajen pristop v tej smeri je razdelitev fonema na manjše enote, ki so lahko odvisne od predhodnih in naslednjih fonemov v zaporedju. Glede na to odvisnost se fonemi razdelijo na enega, dva ali tri dele; označimo jih kot kontekstno odvisne fonemske podkategorije. Fonemi, ki jih ne delimo na manjše enote, pa označimo kot kontekstno neodvisne – t. i. monofone.

Vpliv koartikulacije je ob uporabi kontekstno odvisnih fonemov upoštevan, vendar za ceno večjega števila elementarnih govornih enot. Zato je ta pristop uporabljen le v prvem sistemu. V drugem, bistveno enostavnejšem, je bila eksperimentalno zanimivejša uporaba monofonov.

3.3 Govorni zbirki SLO-GO in ŠTEVKE

Ker v slovenskem jeziku nimamo splošnejše govorne zbirke, smo se odločili, da preliminarne preizkuse izvedemo na lastni govorni zbirki, ki smo jo zasnovali v sodelovanju s podjetjem Amebis. Zbirka trenutno šteje 45 posnetkov devetih govorcev (sedem moških in dve ženski), ki so posneli vsak po pet povedi. V celoti smo fonemsko označili 15 posnetkov.

Povedi so fonemsko primerno bogate in vsebujejo tudi nenavadne govorne povezave fonemskih zaporedij. Pomembno je, da vsebujejo podobne besede, ki so zelo lahko zamenljive s strani SRG. Vsebina prvih petih povedi, katerih izgovarjave so trenutno v zbirki, se nahaja v tabeli 1.

Tabela 1: Vsebina prvih petih povedi govorne zbirke SLO-GO

Št.	Poved
1	Rdeč suh grm sta stric in teta nesla v gost gozd.
2	Nato je šel kmet po koš in kos kruha.
3	Obme je dal snop na rob tnala.
4	Micka je šla v Anhovo po travo za kravo.
5	A nato je šla pa Ana gor na goro.

Zavedamo se relativne majhnosti zbirke, vendar menimo, da za preliminarne raziskave zadostuje. Za preveritev smo izvedli tudi vzporedni test na lastni obsežnejši govorni zbirki ŠTEVKE [7], ki pa je manj splošna. Vsekakor v prihodnje želimo zbirko SLO-GO razširiti z novimi posnetki. Njena bistvena prednost je, da je posneta v sodobnejših okoljih in je tako bliže dejanskim pogojem delovanja SRG.

3.4 Implementacija besednega razpoznavalnika CSLU-SRG

V skladu z opisanim konceptom in s pomočjo orodja CSLU Speech Toolkit [1] je implementiran testni razpoznavalnik z omejenim slovarjem – CSLU-SRG. Nevronska mreža (MLP) v njem opravlja vlogo akustičnega modela – razvršča vektorje značilk iz posameznih okvirjev v razrede, ki ustrezajo govornim enotam (fonemom). Rezultat tega procesa je verjetnostna matrika, ki se nato s postopkom Viterbijevega iskanja transformira v verjetnosti osnovnih enot razpoznave – besed. Ta postopek iskanja opravlja funkcijo časovnega modela. Pri implementaciji so bile uporabljene kontekstno odvisne fonemske podkategorije.

Podrobnejši opis koncepta nevronskih mrež in implementacije sistema v okolju CSLU se nahaja v [7].

3.5 Implementacija nevronskega razpoznavalnika fonemov NN-SRG

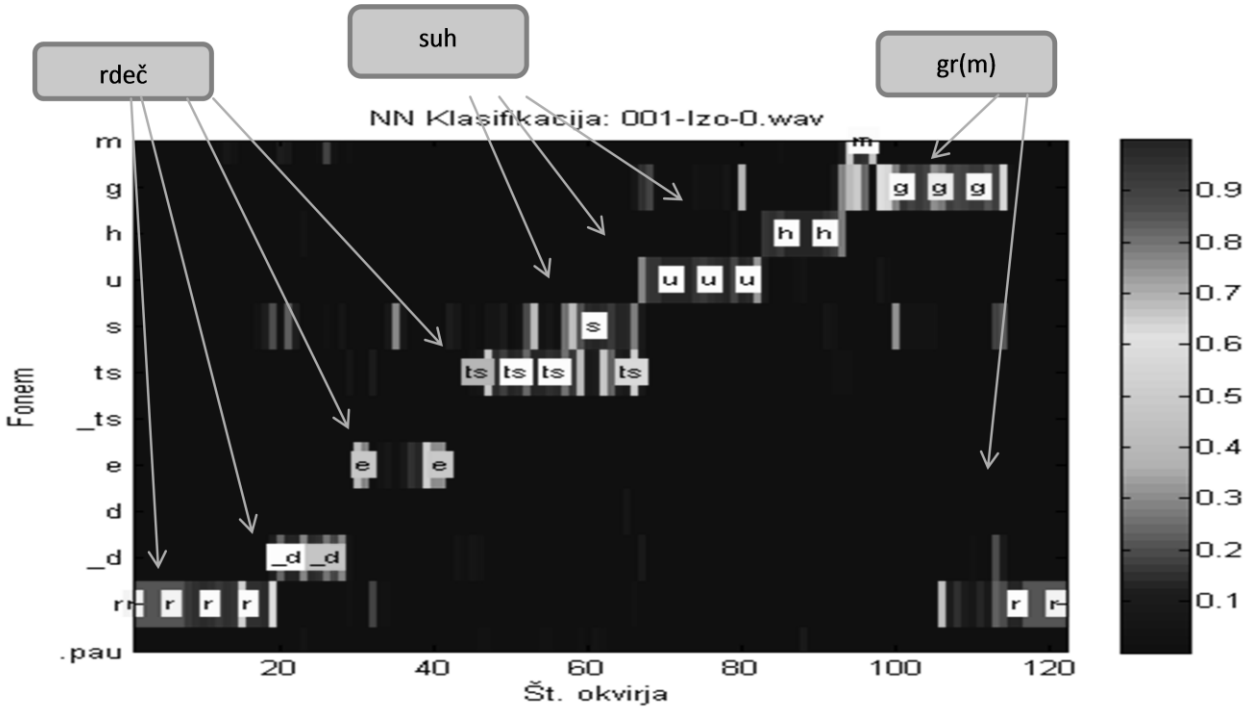
Za lažje proučevanje delovanja in eksperimentiranje s sodobnejšimi zasnovami smo implementirali še enostavnejši razpoznavalnik – NN-SRG. Zasnovan je na nevronski mreži, ki je zelo podobna kot v CSLU-SRG. V opisanem preizkusu smo vse foneme obravnavali kot kontekstno neodvisne celote – monofone.

Glavna razlika v primerjavi s CSLU-SRG je ta, da ni višje ravni združevanja razpoznav fonemov v besede, ampak so rezultat razpoznavanja kar izhodi nevronske mreže; za vsak vektor značilk tako poznamo verjetnostno porazdelitev pripadnosti razredom vseh fonemov v slovarju. Za celotno zaporedje vektorjev značilk dobimo t. i. klasifikacijsko oziroma verjetnostno matriko – podobno kot pri CSLU-SRG.

Izračunano verjetnostno predstavitev lahko koristno uporabimo tudi pri razpoznavanju kompleksnejših govornih enot na višji ravni – npr. besed. Slika 1 prikazuje primer preproste sestave zaporedja razpoznanih besed iz verjetnostne matrike. Na tej podlagi lahko uporabimo različne metode časovnega združevanja teh delnih razpoznav v daljše govorne enote – besede, fraze, povedi …

4 REZULTATI

Osnovni namen opisanih preliminarnih preizkusov je dobiti boljši vpogled v uspešnost delovanja enostavnejše zasnove SRG. Rezultati v tem primeru niti niso najpomembnejši, ker bi bilo iluzorno pričakovati boljše rezultate od trenutno najbolj dodelanih konceptov, ki so že dolga leta v uporabi in so predmet nenehnih testiranj in izboljšav. Kljub temu v predstavljenih rezultatih najdemo nekaj zanimivih spoznanj.



Slika 1: Prikaz uporabe klasifikacije nevronske mreže (NN-SRG) za razvrstitev v razrede govornih enot – fonemov in zaporedja besed "rdeč suh grm" na prvem delu posnetka "001-Izo-0.wav"

4.1 CSLU-SRG

CSLU-SRG smo preizkusili na govorni zbirki SLO-GO; v učni množici je bilo 21, v razvojni 10 in v testni devet posnetkov. Zaradi manjšega števila fonemsko označenih posnetkov je učenje sistema potekalo v dveh korakih. V prvem se je v učni množici uporabila samo fonemsko označena podmnožica (osem posnetkov). Nato je tako naučen sistem avtomatsko fonemsko označil še preostali del učne množice. V drugi ponovitvi učenja je tako bilo v učni množici vseh 21 posnetkov. Uspešnost posameznih iteracij procesa učenja se je preverila na razvojni množici (10 posnetkov); najuspešnejša iteracija na razvojni množici se je uporabila za končni preizkus na testni množici. Rezultat razpoznavanja na testni množici je prikazan v tabeli 2. V prvi vrstici sredinskega stolpca je najprej zapisana pravilna poved, pod njo pa še razpoznana vsebina. Uspešnost na ravni besed je bila 88-odstotna in 44-odstotna (štiri od devetih) na ravni povedi (oznaka +). Če napake upoštevamo manj strogo, potem so skoraj v celoti pravilne še tri povedi (oznaka +-, napake so izključno pri manj pomembnih besedah); še več, tudi v preostalih dveh povedih z napako (oznaka -) so ključne besede večinoma uspešno razpoznane. Ugotovimo lahko, da sama numerična uspešnost ne pove vsega; vpliv napak na samo uspešnost razpoznave je relativen in ga z ustrezno višjo ravnjo obdelave lahko bistveno zmanjšamo ali celo odpravimo. Pomembno je le, da poleg trenutno najboljše ohranimo tudi nekaj manj uspešnih hipotez.

Tabela 2: Prikaz uspešnosti besednega razpoznavalnika CSLU-SRG na testni množici govorne zbirke SLO-GO

	Pravilne in razpoznane povedi (zap. št., vsebina, uspešnost)	
1	rdeč suh grm sta stric in teta nesla v gost gozd rdeč suh grm sta stric in teta nesla v gost gozd	+
2	nato je šel kmet po koš in kos kruha nato je šel kmet po koš in kos kruha	+
3	nato je šel kmet po koš in kos kruha nato je šel kmet po koš in kos kruha	+
4	obme JE DAL snop na rob tnala obme NA ### snop na rob tnala	-
5	obme JE dal snop na rob tnala obme ## dal snop na rob tnala	+-
6	# micka je sla v ANHOVO po TRAVO ZA kravo V micka je sla v A po PO NA kravo	-
7	micka je šla v anhovo po travo za kravo # micka je šla v anhovo po travo za kravo V	+-
8	a nato je šla pa ana # gor na goro a nato je šla pa ana V gor na goro	+-
9	a nato je šla pa ana gor na goro a nato je šla pa ana gor na goro	+

Za lažjo primerjavo z NN-SRG naj omenimo še uspešnost sistema na učni množici na ravni okvirjev; ta je na učni množici znašala največ 89 odstotkov v 169 kontekstno odvisnih fonemskih podkategorijah.

4.2 NN-SRG - nevronski razpoznavalnik fonemov

V primerjavi s sistemom CSLU-SRG so rezultati NN-SRG pričakovano slabši, saj je razpoznava na ravni posameznih okvirjev praviloma slabša od razpoznave na višji ravni fonemov oziroma besed (CSLU-SRG). Prav tako je treba poudariti, da NN-SRG uporablja

kontekstno neodvisne (celovite) foneme, ki jih je bistveno manj kot kontekstno odvisnih fonemskih kategorij v CSLU-SRG. V tabeli 3 je prikazana uspešnost na govorni zbirki SLO-GO in v tabeli 4 še uspešnost na zbirki ŠTEVKE.

Rezultati pokažejo, da se oba sistema bistveno bolje prilagodita učni množici, vendar je zares pomembna uspešnost na testni množici, ki jasno izraža splošnost naučenega znanja. Pri rezultatih na testni množici so podane tri uspešnosti; oznaka »top« pomeni uvrstitev pravilnega rezultata na lestvici dejanskih razpoznav, in sicer:

- »top1« pomeni prvo mesto,
- »top2« prvi dve in
- »top3« prva tri mesta.

Tabela 3: Uspešnost na ravni okvirjev (FSR[1]) razpoznavalnika NN-SRG v odvisnosti od števila nevronov v skriti plasti (N) – govorna zbirka SLO-GO

N	Učna množica [%]	Testna množica »top1« , »top2« , »top3« [%]
20	66.6	47.4 , 58.1 , 62.1
50	85.4	57.5 , 65.8 , 71.9
100	83.7	54.9 , 67.6 , 74.5
150	82.5	54.5 , 63.4 , 68.8

Tabela 4: Uspešnost na ravni okvirjev (FSR) razpoznavalnika NN-SRG v odvisnosti od števila nevronov v skriti plasti (N) – govorna zbirka ŠTEVKE

N	Učna množica [%]	Testna množica »top1«, »top2«, »top3« [%]
100	75.8	64.3 , 76.4 , 81.8
150	77.5	65.0 , 76.5 , 81.6
200	79.5	65.6 , 77.2 , 82.0
300	83.0	65.0 , 77.2 , 82.5

V tabeli 3 vidimo, da je posploševanje naučenega znanja manj izrazito – to pripišemo manjšemu številu učnih primerov. V obeh tabelah pa opazimo, da je pogosto pravilni rezultat na drugem ali tretjem mestu pri dejanski razpoznavi. Zato je treba ohraniti vsaj nekaj prvih hipotez tudi na višjih ravneh razpoznavanja, sicer razpoznava ne more biti več pravilna.

5 SKLEP

Opozorili smo na nekaj glavnih slabosti SRG v primerjavi z novejšimi dosežki razvoja na področju razpoznavanja govora in sorodnih področij. Kot pomembno alternativo smo prikazali enostavnejšo zasnovo SRG, ki ima bolj transparentno razdelitev na več plasti razpoznavanja. Taka zasnova se nam zdi primernejša za uporabo novejših tehnik in pristopov ter seveda izrabo vzporednih in porazdeljenih načinov procesiranja. Seveda pa je še v začetni fazi razvoja.

Glavni namen realizacije obeh eksperimentalnih sistemov se ne konča z opisanimi preizkusi. Ti so preliminarne narave in pomenijo začetek raziskav na področju sodobnejše zasnove enostavnejših SRG. V prihodnosti nameravamo predvsem NN-SRG razširiti in preizkusiti z novimi idejami in pristopi k snovanju sodobnih SRG. Prav tako želimo razširiti govorno zbirko SLO-GO s čim več realnimi posnetki. Predvsem pa nameravamo raziskati možnosti nadomestitve vgrajenega znanja v tovrstnih sistemih z avtomatsko pridobljenim iz obstoječih podatkov in preveriti optimalnost procesa parametrizacije. Obstoj jezikovnih tehnologij je namreč ključnega pomena tudi za obstoj slovenskega jezika.

[1] FSR ali »Frame Success Rate«.

LITERATURA

[1] CSLU Speech Toolkit, http://www.cslu.ogi.edu/toolkit/.

[2] Victor Zue, James Glass, Michael Phillips, and Stephanie Seneff. 1989. The MIT SUMMIT Speech Recognition system: a progress report. In *Proc. of the workshop on Speech and Natural Language* (HLT '89), pp.179–189.

[3] Hu, Zhihong and Schalkwyk, Johan and Barnard, Etienne and Cole, Ronald, "Speech Recognition Using Syllable-Like Units", in *Proc. of International Conference on Spoken Language Processing(ICSLP)*, pp. 1117–1120, Acoustic Soc. of America, Philadelphia, USA, Oct, 1996.

[4] Ladan Golipour, D. O'Shaughnessy, A segmental non-parametric-based phoneme recognition approach at the acoustical level, Computer Speech & Language, Vol. 26, Issue 4, August 2012, Pages 244–259, ISSN 0885-2308.

[5] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, Deep Neural Networks for Acoustic Modeling in Speech Recognition, IEEE Signal Processing Magazine, **29**, November 2012.

[6] ROZMAN, Robert, KODEK, Dušan. Using asymmetric windows in automatic speech recognition. *Speech communication* [Print ed.], 2007, vol. 49, no. 4, str. [268]–276.

[7] ROZMAN, Robert. Nesimetrične okenske funkcije v sistemih za razpoznavanje govora : doktorska disertacija. Ljubljana: [R. Rozman], 2005. VI, 128 f., ilustr.

[8] ROZMAN, Robert, KODEK, Dušan. Improving speech recognition robustness using non-standard windows. *The IEEE Region 8 EUROCON 2003 : computer as a tool :* IEEE, cop. 2003, vol. 2, str. 171–174.

[9] ROZMAN, Robert, ŠTRANCAR, Andrej, KODEK, Dušan. Povečevanje robustnosti sistemov za razpoznavanje govora in optimizacija procesa parametrizacije. V: ZAJC, Baldomir (ur.). *Zbornik desete Elektrotehniške in računalniške konference ERK 2001*, zv. B, str. 257–260.

Robert Rozman je leta 2005 doktoriral s področja sistemov za razpoznavanje govora na Fakulteti za računalništvo in informatiko v Ljubljani, kjer je trenutno višji predavatelj. Njegovo raziskovalno področje zajema računalniške arhitekture, avtomatizirana bivalna okolja, digitalno procesiranje signalov in razpoznavanje govora.