Govorna baza ŠTEVKE in raziskave robustnosti sistemov za razpoznavanje govora

Robert Rozman, Dušan M. Kodek

Laboratorij za arhitekturo in procesiranje signalov
Fakulteta za računalništvo in informatiko
Univerza v Ljubljani
Tržaška 25, 1001 Ljubljana, Slovenija
{rozman, kodek}@fri.uni-lj.si

Povzetek

V osnovnih orisih predstavljamo področja raziskovalnega dela v naši skupini. Najprej opisujemo govorno bazo ŠTEVKE, ki je bila osnova pri oblikovanju opisanih sistemov za razpoznavanje tako ločenih besed kot zaporedij števk. Posebna pozornost je posvečena povečevanju robustnosti tovrstnih sistemov in s tem tudi večje uspešnosti v realnih pogojih delovanja. Pri tem smo poskušali koristno uporabiti nekatera spoznanja o človekovem slušnem zaznavanju, še posebej tista, ki so tesno povezana s človekovo robustnostjo pri razpoznavanju govora; ta je zaenkrat za avtomatske sisteme še nedosegljiva. Podani so tudi rezultati praktičnih meritev, ki potrjujejo tezo, da si lahko s spoznanji o lastnostih človekovega sluha precej pomagamo pri oblikovanju robustnejših sistemov za razpoznavanje govora.

Abstract

The basic areas of research work in our group is presented. First, speech corpora ŠTEVKE is described. It served as base for recognition systems of isolated words and connected digits. Special attention is dedicated to increasing of this type systems' robustness and as consequence also higher rate of success in real working conditions. Doing that, we tried to use knowledge on some findings about human auditory perception, specially those tightly connected to human's speech recognition robustness that is still unreachable for automatic systems. Practical results are given which support the thesis that the knowledge of human hearing can be successfully used for design of more robust speech recognition systems.

1. Uvod

Pri razvoju sistemov za razpoznavanje govora (v nadaljevanju SRG) so ustrezne govorne baze nepogrešljive. V slovenskem prostoru so te še posebej redke, kar lahko predstavlja dokaj resno oviro pri raziskovalnem delu na tem področju. V nadaljevanju je predstavljena govorna baza ŠTEVKE. V njej se nahajajo izgovorjave večjega števila govorcev iz vse Slovenije. Na njeni osnovi smo s pomočjo nekaterih programskih orodij razvili več razpoznavalnikov z majhnim naborom besed oziroma slovarjem. Dva med njimi v nadaljevanju tudi opisujemo. Prvi razpoznava ločeno izgovorjene besede. Kot tak v novejšem času sicer ni posebej zanimiv za praktično uporabo, je pa zaradi enostavne zasnove zelo primeren za nekatere temeljne raziskave robustnosti[1], ki smo jih in jih še opravljamo. Drugi razpoznavalnik, ki nastaja na osnovi govorne baze ŠTEVKE, pa je razpoznavalnik poljubno dolgega zaporedja števk ("0" do "9"). V prihodnje nameravamo bazo ŠTEVKE še dopolniti z bolj spontanimi izgovorjavami; prav spontanost je namreč eden tistih pomembnih atributov sodobnih govornih baz, ki ŠTEVKAM še najbolj primanjkuje.

V 2. poglavju je opisana sama govorna baza, njene posebnosti ter postopek testiranja robustnosti SRG. V nadaljevanju sta podrobneje predstavljena še dva sistema za razpoznavanje govora in v povezavi s prvim še nekatere izboljšave ter rezultati praktičnih testiranj njegove uspešnosti oziroma robustnosti. V zaključku povzemamo ugotovitve in podajamo nekaj pogledov v prihodnost.

2. Govorna baza ŠTEVKE

2.1. Zajemanje in priprava

Govorna baza je bila zajeta[2] v letih 1992-1993 v - takrat še - Mikroračunalniškem laboratoriju Fakultete za elektrotehniko in računalništvo. V bazi se nahajajo izgovorjave 780 govorcev iz vse Slovenije (skupna dolžina posnetkov je slabih 5 ur), posnete preko direktno vzpostavljenih telefonskih linij. Vsak izbran govorec je v naključnem vrstnem redu izgovoril 13 besed (števke od "0" do "9" ter besede "ja", "ne" in "stop"). Postopek je vodil operater, ki je večje napake tudi sproti odpravljal.

V govorni bazi so dokaj sorazmerno zastopani govorci obeh spolov, vseh starosti, stopenj izobrazbe in večjih naročnih skupin. Podrobnejši podatki o postopku zajemanja, sestavi množice govorcev in drugih značilnostih baze se nahajajo v literaturi (Kodek s sodelavci, 1994; Majnik, 1993).

Postopku snemanja je sledilo označevanje izgovorjenih besed – labeliranje. Vsaka beseda je bila praviloma posneta tako, da se je pred in za njo nahajal še signal tišine oziroma okolja. Za vsako besedo sta bila "ročno" označena njen začetek in konec.

Po labeliranju je sledil ponovni vsebinski pregled baze. Nekateri govorci so bili razporejeni v skupine z opombami. To pomeni, da ima njihov posnetek neko posebnost, ki je lahko vezana na motnje, napake pri zajemu ali govorčeve posebnosti oziroma njegovo večje odstopanje pri izgovorjavi. Ostali nerazporejeni govorci predstavljajo "zlato sredino", ki vsebuje kar precej

[1] Lastnost sistema, da je uspešen tudi v zmerno spremenjenih pogojih oziroma ob prisotnosti motenj.

[2] Frekvenca vzorčenja je bila 8 kHz.

variabilnosti, ki jo v procesu učenja sistem poskuša kar najbolje obvladati.

Zaradi novejših potreb pri snovanju razpoznavalnikov, ki za temeljno enoto razpoznavanja uporabljajo foneme, smo v izgovorjavah pri 100 naključno izbranih govorcih dodatno označili še meje med temi podbesednimi enotami.

2.2. Posebnosti

Glavna posebnost govorne baze ŠTEVKE je večje število govorcev (780) z vseh področij Slovenije. To pomeni, da je v izgovorjavah dokaj dobro zajeta tudi vsa raznolikost našega jezikovnega prostora, še posebej raznolikost narečnih skupin. Obenem lahko ugotovimo, da je baza dokaj reprezentativen vzorec množice potencialnih govorcev tudi po ostalih atributih (spol, starost in izobrazba).

Prav zaradi omenjenih posebnosti in raznovrstnih pogojev snemanja[3] je baza poleg razpoznavanja zanimiva tudi za raziskave na drugih področjih govornih tehnologij, kot so npr. razpoznavanje govorcev, vpliv telefonskega kanala na prenašan govorni signal in še nekatere druge.

Med osnovne omejitve baze ŠTEVKE sodi majhen slovar (13 besed) in relativna nespontanost govorcev[4]. Te pomanjkljivosti bomo poskušali v prihodnosti odpraviti z ustreznimi dopolnitvami, ki jih bomo posneli ob delovanju eksperimentalnega govornega vmesnika.

2.3. Testiranje robustnosti SRG

Kar precej sistemov za razpoznavanje govora ima pri vključitvi v realno okolje delovanja nemalo problemov s svojo uspešnostjo. Njihov razvoj poteka na osnovi govornih baz, ki pa so le boljši ali slabši "posnetek" realnih razmer. V primeru izrazitega ujemanja govorne baze z realnimi pogoji je teh problemov precej manj. Seveda pa to pomeni kar precejšnjo omejitev (SRG dosegajo načrtovano uspešnost le v enakih pogojih), kar lahko povzroči potrebo po specifičnih govornih bazah za različne aplikacije. Zato je vse bolj pomembna lastnost sodobnih SRG poleg uspešnosti tudi robustnost, ki zagotavlja primerno uspešnost tudi v drugačnih pogojih, ki niso bili izrazito prisotni v govorni bazi. Na ta način lahko prihranimo snemanje specifične govorne baze, kar je običajno dolgotrajen in drag proces.

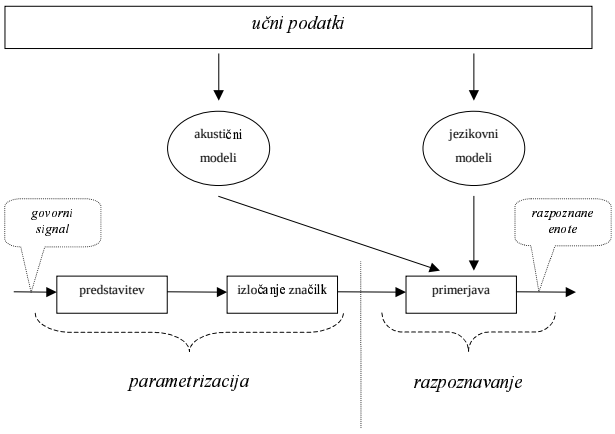
Pri oblikovanju SRG nas je zato zanimala tudi njihova robustnost in možnosti za njeno izboljšavo. Robustnost sistema smo ocenjevali tako, da smo preverjali njegovo uspešnost tudi na izgovorjavah z dodanimi aditivnimi zvočnimi motnjami[5] in konvolutivnimi popačitvami[6]. Tako smo uspešnost sistemov merili na treh oziroma v enem primeru štirih skupinah testnih množic[7]:

- "Čist signal" - nespremenjena testna množica.
- "Aditivne motnje" - 6 testnih množic, ki smo jih dobili, ko smo osnovni testni množici dodajali naslednje aditivne motnje[8]:
  - govor skupine ljudi v ozadju,
  - hrup v pilotski kabini bojnega letala,
  - hrup v tovarni,
  - hrup v osebnem avtomobilu,
  - roza šum,
  - beli šum.
- "Aditivne motnje + konvolutivni vpliv" – 6 testnih množic, ki smo jih dobili z dodatnim nizkoprepustnim[9] filtriranjem testnih množic iz prejšnje skupine.
- "Pasovne motnje" – 4 testne množice, ki smo jih dobili z dodajanjem frekvenčno omejenega (800-1000Hz) signala belega šuma izhodiščni testni množici z različnimi ciljnimi razmerji energij SNR, in sicer 12, 6, 0 in –6 dB.

V učni množici se je nahajalo 200 govorcev, v testni in validacijski pa po 100 govorcev, ki niso bili hkrati tudi v učni množici. V osnovnem sistemu so bile uporabljene standardne MFCC značilke z dodano energijo okvirja in delta značilkami prvega reda – skupaj 26 značilk za vsak okvir. Uspešnost sistema je v nadaljevanju izražena z odstotkom pravilno razpoznanih besed.

3. Razpoznavanje ločenih besed

Razpoznavalnik je bil razvit na osnovi govorne baze ŠTEVKE in programskega orodja HTK. Zgradba razpoznavalnikov ločenih besed je enostavna in precej splošna (slika 1). V fazi parametrizacije izračunamo kompaktno predstavitev posameznega dela govornega signala – t.i. vektor značilk, katerih zaporedje vstopa v fazo razpoznavanja, kjer izračunamo podobnost med vhodnim zaporedjem vektorjev značilk z že vnaprej shranjenimi opisi oziroma modeli posameznih govornih enot in jezika. V najenostavnejšem primeru razpoznavanja ločenih besed kot končni izid izberemo besedo, kateri ustreza akustični model z najvišjo mero podobnosti. To smo storili tudi v primeru opisanega razpoznavalnika.



Slika 1: Zgradba splošnega SRG.

V novejšem času smo prav zaradi enostavnosti sistema na njem izvedli kar nekaj raziskav na področju uvajanja nekaterih lastnosti človekovega slušnega zaznavanja z

[3] Telefonska zveza je bila vzpostavljena z vsakim govorcem posebej na njegovem domu.

[4] O snemanju obveščeni govorci običajno govorijo drugače kot sicer oziroma večinoma bolj razločno.

[5] Sem štejemo motnje, ki običajno nastanejo v okolju govorca ali pri samem prenosu in se signalu prištevajo.

[6] Sem štejemo konvolutiven vpliv elementov, ki sestavljajo prenosno pot govornega signala od govorca k poslušalcu.

[7] Uspešnost na določeni skupini množic je povprečje uspešnosti na vseh množicah v skupini (razen pri "Pasovnih motnjah").

[8] Uporabljeni so bili signali motenj iz govorne baze NOISEX. Ciljno razmerje SNR je bilo 12 dB.

[9] Nizkoprepustni filter predstavlja primer konvolutivnega vpliva, ki lahko nastane kot posledica npr. vpliva telefonskega kanala, karakteristik mikrofona, ovire v akustičnem okolju in podobno.

željo, da jih preizkusimo v praksi in z njimi povečamo njegovo robustnost. V nadaljevanju poglavja so predstavljeni izsledki raziskav skupaj z rezultati meritev robustnosti dobljenih sistemov.

3.1. Uvajanje izbranih lastnosti človekovega slušnega zaznavanja

V želji po večji robustnosti sistema smo se odločili v sistem vključiti nekaj tehnik, ki implementirajo nekatere zanimive lastnosti človekovega sluha. Med njimi smo izbrali tiste, ki so izražale:

- povezanost z robustnostjo človekovega sluha,
- možnost enostavne, časovno nezahtevne implementacije.

Obstoječi osnovni sistem smo modificirali v treh določenih detajlih tako v fazi parametrizacije kot tudi v fazi razpoznavanja. V nadaljevanju so te modifikacije v osnovnih značilnostih tudi opisane, več podrobnosti pa se nahaja v literaturi (Rozman, 1999).

3.1.1. Nesimetrične okenske funkcije

Kar nekaj raziskav potrjuje dejstvo, da je človekov sluh neobčutljiv na zmerne fazne popačitve govornega signala. Vsekakor je ta lastnost povezana s človekovo robustnostjo. Zato smo v fazi parametrizacije najbolj pogosto uporabljano okensko funkcijo (t.i. Hammingovo okno) nadomestili s tistimi, ki imajo na račun sicer večjih faznih popačitev[10], ki jih vnašajo v govorni signal, boljše druge lastnosti – predvsem amplitudni odziv.

Praktično uspešnost tako izboljšanih SRG oziroma robustnost smo izmerili na prej opisanih skupinah testnih množic. Naj opozorimo, da je bil sistem vselej naučen zgolj na obstoječi učni množici, brez kakršnihkoli dodatnih motenj. Za primerjavo robustnosti v tabeli 1 smo izbrali osnovni sistem z dodanim RASTA filtrom (opisan v 3.1.2), ki uporablja v fazi parametrizacije Hammingovo okno, in enak sistem z nesimetričnim oknom, ki je med vsemi preizkušenimi pokazal največjo robustnost.

Tabela 1: Izboljšava robustnosti osnovnega sistema z dodanim RASTA filtrom pri uporabi nesimetričnega okna.

	Hammingovo okno	Nesimetrično okno
Čist signal	97,77	97,54
Aditivne motnje	75,69	76,39
Aditivne motnje + konv. vpliv	43,32	68,14

3.1.2. Transformacije signalov značilk

Praktične meritve lastnosti človekovega sluha pokažejo na kar nekaj zelo zanimivih lastnosti; predvsem sta pomembni naslednji:

- človekova slušna zaznava ni odvisna le od trenutnega zvočnega dogodka, ampak tudi od dogodkov neposredno pred in po njem (tej lastnosti rečemo tudi časovno maskiranje),
- človekov sluh je najbolj občutljiv le na tiste spremembe v kratkočasovnem spektru govornega signala, ki potekajo v določenem razponu hitrosti[11].

Opisani lastnosti je mogoče enostavno in časovno nezahtevno implementirati tudi v SRG. Časovno maskiranje lahko posnemamo s preprostim modelom, medtem ko človekovo občutljivost na spremembe v spektru signala najbolje ponazorimo s pasovno prepustnim filtrom (t.i. RASTA filter). Delno pa to storimo tudi z normalizacijo[12] signalov[13] posameznih značilk.

Na podlagi meritev robustnosti (tabela 2) lahko ugotovimo, da uporaba vsakega med prej omenjenimi konceptoma poveča robustnost osnovnega sistema. Pri tem je potrebno dodati še to, da lahko model časovnega maskiranja in RASTA filter uporabimo za delo v realnem času. Postopek normalizacije signalov značilk, ki je v tem primeru najbolj povečal robustnost sistema, je pri izračunu upošteval vse vrednosti signala v izgovorjavi. Če bi normalizacijo želeli opraviti z manjšo zakasnitvijo oziroma v realnem času, bi morali upoštevati precej krajši odsek signala značilke, kar bi vsekakor zmanjšalo v tabeli 2 prikazano robustnost te metode.

Tabela 2: Izboljšava robustnosti osnovnega sistema z dodatnimi transformacijami signalov značilk.

	Čist signal	Aditivne motnje	Aditivne motnje + konvolutivni vpliv
Osnovni sistem	97,77	26,59	8,55
+ model časovnega maskiranja	97,46	65,23	21,14
+ RASTA filter	97,77	75,69	43,32
+ normalizacija	97,77	80,49	51,51

3.1.3. Večtokovna in večpasovna paradigma

Nekaj že zelo zgodnjih raziskav na področju psihoakustike (Fletcher, 1953) kaže na to, da človekov sluh opravlja razpoznavanje v različnih frekvenčnih pasovih povsem neodvisno (t.i. večpasovna paradigma). S tem spoznanjem si je mogoče razložiti človekovo uspešnost razpoznavanja govora tudi ob izrazitejši prisotnosti frekvenčno omejenih motenj, ki pa so za uspešnost večine SRG kljub svoji omejenosti "usodne". To potrjujejo tudi rezultati osnovnega sistema z normalizacijo signalov značilk v tabeli 3.

V zadnjem času pa se vse bolj uveljavlja tudi posplošitev omenjene teze – t.i. večtokovna paradigma (slika 2). Ta pomeni, da v sistemu deluje več (lahko po

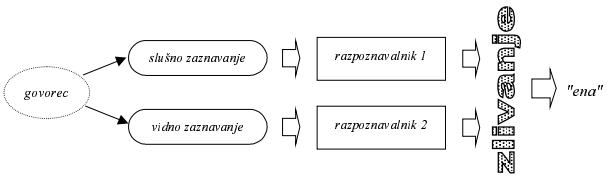
[10] Po zgledu lastnosti človekovega sluha smo fazne popačitve pri načrtovanju okna zanemarili.

[11] Vrh občutljivosti doseže pri spremembah spektra, ki potekajo s frekvencami med 2-10 Hz.

[12] Normaliziramo povprečno vrednost in varianco signala značilke v celotni izgovorjavi.

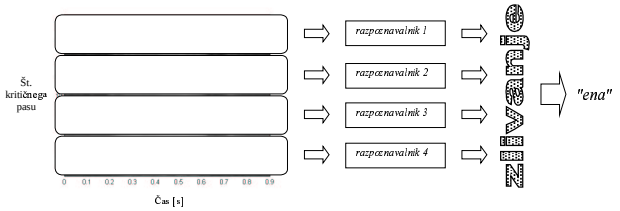
[13] Za signal značilke v splošnem imenujemo časovno diskretni signal, ki ga dobimo s predstavitvijo časovnega poteka vrednosti značilke.

zasnovi tudi različnih) razpoznavalnikov na večjem številu različnih informacijskih tokov, ki so lahko tudi različnih izvorov – npr. zvočni, slikovni. Na koncu rezultate vseh delnih razpoznav združimo v enotni izid razpoznavanja – temu postopku rečemo zlivanje. Praktični poskusi pokažejo, da je tak sistem uspešen in v pogojih frekvenčno omejenih motenj robusten že pri uporabi enostavnih metod zlivanja, kot sta npr. seštevanje delnih izidov ali "večinski" sistem.



Slika 2: "Večtokovni" SRG.

Zaradi obstoječih omejitev[14] smo realizirali večpasovni sistem (slika 3) tako, da smo večpasovni pristop uveljavili na osnovnem sistemu z normalizacijo značilk, ki je po rezultatih v tabeli 2 najbolj robusten.



Slika 3: Implementacija "večpasovnega" sistema.

Na podlagi vseh rezultatov v tabeli 3 lahko ugotovimo, da večpasovni pristop zelo poveča neobčutljivost na frekvenčno omejene motnje – ta je že dokaj primerljiva s tisto pri človeku. Hkrati pa tudi vidimo (tabela 4), da nov pristop zmanjša že doseženo robustnost osnovnega sistema z normalizacijo v primeru aditivnih motenj in konvolutivnih vplivov. Temu dejstvu kaže v prihodnje posvetiti več pozornosti.

4. Razpoznavanje zaporedij števk

Pri oblikovanju razpoznavalnika smo uporabili govorno bazo ŠTEVKE v originalni obliki (celotne izgovorjave posameznih govorcev). Osredotočili smo se na kompaktno obliko razpoznavalnika, ki temelji na uporabi nevronske mreže kot klasifikatorja fonemskih enot in Viterbijevem iskalnem algoritmu. Tovrstna realizacija je prostorsko relativno nezahtevna in je zato primerna za realizacijo tudi v aparaturnih sistemih.

Sistem smo naučili na delnih izgovorjavah 468 govorcev in preizkusili na testni množici, v kateri se nahaja 156 govorcev, ki niso v učni oziroma validacijski množici. Sistem pravilno razpozna 96.55% besed, kar je za sistem v razvoju dokaj vzpodbuden rezultat.

5. Zaključek

Predstavljeni so bili rezultati nekaterih raziskav na področju razpoznavanja govora z majhnim slovarjem. V primeru razpoznavalnika ločenih besed so podane nekatere izboljšave sistema, ki pripomorejo k večji robustnosti v realnih pogojih. Rezultati so potrdili dejstvo, da si lahko z uporabo določenih manj zahtevnih lastnosti človekovega sluha precej pomagamo tudi pri oblikovanju "človeku" podobnih SRG – predvsem v pozitivnih lastnostih, ki jih človekov sluh ima. Opisan je bil tudi razpoznavalnik nizov števk, ki je še v razvojni fazi, a vendar že kaže obetavne rezultate.

Tabela 3: Primerjava robustnosti "konvencionalnega" in "večpasovnega" pristopa pri osnovnem sistemu z normalizacijo v primeru pasovnih motenj.

	Osnovni sistem + normalizacija	"Večpasovni" osnovni sistem + normalizacija
Čist signal	97,77	90,38
SNR 12 dB	76,46	89,08
SNR 6 dB	62,62	88,69
SNR 0 dB	37,38	87,00
SNR –6 dB	19,46	80,31

Tabela 4: Primerjava robustnosti "večpasovnega" in "konvencionalnega" pristopa pri osnovnem sistemu z normalizacijo v primeru frekvenčno neomejenih motenj.

	Osnovni sistem + normalizacija	"Večpasovni" osnovni sistem + normalizacija
Čist signal	97,77	90,38
Aditivne motnje	80,49	56,27
Aditivne motnje + konvolutivni vpliv	51,51	45,87

V kratkem želimo realizirati avtomatski govorni odzivnik, kjer bomo opisane razpoznavalnike uporabili, obenem pa bomo s sprotnim snemanjem dialogov lahko pripravili tudi ustrezne dopolnitve k obstoječi govorni bazi, ki bo tako boljša osnova za nadaljnje raziskave. Nadaljevali bomo tudi s proučevanjem robustnosti SRG in njihovih kompaktnih, časovno manj zahtevnih oziroma kompleksnih realizacij.

6. Literatura

Harry Fletcher. 1953. *Speech and Hearing in Communication.* New York: Krieger.

Dušan Kodek s sodelavci. 1994. *Razvoj in izdelava sistema za razpoznavanje izoliranih besed slovenskega govora.* Končno poročilo, Fakulteta za elektrotehniko in računalništvo, Univerza v Ljubljani.

T. Majnik. 1993. *Analiza baze podatkov telefonskega govora.* Diplomska naloga, Fakulteta za elektrotehniko in računalništvo, Univerza v Ljubljani.

R. Rozman. 1999. *Uporaba spoznanj o človekovi slušni percepciji v sistemu za razpoznavanje govora.* Magistrsko delo, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani.

[14] Na voljo je le govorna baza z zvočnimi posnetki.