RAZVOJ ZBIRKE SLOVENSKEGA ČUSTVENEGA GOVORA IZ RADIJSKIH IGER — EmoLUKS

Tadej JUSTIN, France MIHELIČ
Univerza v Ljubljani, Fakulteta za elektrotehniko

Janez ŽIBERT
Univerza na Primorskem, Fakulteta za matematiko, naravoslovje in informacijske tehnologije

Justin, T., Žibert J., Mihelič F. (2015): Razvoj zbirke slovenskega čustvenega govora iz radijskih iger — EmoLUKS. Slovenščina 2.0, 2015 (2): 1–44.
URL: http://www.trojina.org/slovenscina2.0/arhiv/2015/2/Slo2.0_2015_2_01.pdf.

V prispevku predstavljamo graditev slovenske zbirke čustvenega govora za umetno tvorjenje govora in hkrati raziščemo tudi možnosti njene uporabe pri razpoznavanju čustvenega stanja govorca. V prispevku se osredotočamo na opis razvite metodologije za označevanje paralingvistične informacije v govoru na primeru označevanja čustvenih stanj v slovenskih radijskih igrah. Zbirka vsebuje govorne zvočne signale sedemnajstih radijskih iger. Trenutno označeno gradivo obsega čustven govor enega govorca in ene govorke. Čustvene oznake posnetkov smo pridobili s pomočjo dvostopenjskega označevanja s petimi prostovoljnimi označevalci, ki so označili posnetke v dveh časovno ločenih intervalih. Način označevanja omogoča medsebojno primerjavo oznak označevalcev. S pomočjo označenega gradiva v obeh iteracijah poročamo o konsistentnosti označevalcev in ujemanju njihovih mnenj. Na podlagi večinskega mnenja pridobljenih čustvenih oznak vsakemu posnetku pripišemo tisto čustveno oznako, ki je bila med označevalci največkrat izbrana, in tako označene posnetke združimo v zbirko čustvenega govora EmoLUKS, ki jo kvantitativno in kvalitativno ovrednotimo z uporabo uveljavljenega samodejnega sistema za razpoznavanje čustvenih stanj govorca. Konsistentnost oznak ovrednotimo z dvorazrednim in sedemrazrednim od govorca odvisnim razvrščevalnikom čustvenih stanj. Uspešni rezultati razpoznavanja dodatno potrjujejo, da podatkovna zbirka kljub svoji zahtevnosti vsebuje jasno izražena čustvena stanja govorca.

Ključne besede: zbirka čustvenega govora, razpoznavanje čustvenih stanj govorca

1 UVOD

Samodejni sistemi, namenjeni razpoznavanju in/ali tvorjenju umetnega govora, so močno odvisni od govornih podatkovnih zbirk (Ayadi idr. 2011; Koolagudi in Rao 2012). Ne le dobro zasnovana zbirka, temveč tudi jezik govora igra pomembno vlogo pri razvoju in nadaljnji aplikativni rabi tovrstnih sistemov. Če želimo razviti samodejni razpoznavalnik slovenskega govora ali sistem za umetno tvorjenje slovenskega govora, moramo imeti na voljo dovolj posnetkov govorjenega slovenskega jezika in tudi pripadajoče natančne prepise. Tako urejena zbirka izpolnjuje minimalno zahtevo za uporabo v samodejnih sistemih, namenjenih tvorjenju ali razpoznavanju govora. Natančnost pri samodejnih sistemih za razpoznavanje govora (Juang in Rabiner 1991) in kakovost umetnega govora (Vesnicer in Mihelič 2004) se izboljša, če tovrstne sisteme razvijemo tako, da prepise, ki so del govorne zbirke, pretvorimo s pomočjo pravil glasoslovja v eno od fonetičnih abeced. Na področju govornih tehnologij imenujemo tovrstni samodejni postopek grafemsko-fonemska pretvorba. Postopki, ki omogočajo tovrstno preslikavo, so močno odvisni od zakonitosti govorjene besede v določenem jeziku in ne nazadnje tudi od narečja govorca.

Raziskovalci področja govornih tehnologij želimo razpolagati z govornimi podatkovnimi zbirkami, odvisnimi od jezika, ki zajemajo čim več jezikovnih prvin tako pisnega kot tudi govorjenega jezika. Le dobro zasnovana zbirka omogoča kakovostno tvorbo ali razpoznavanje govora s pomočjo sodobnih, naprednih pristopov pri razvoju tovrstnih samodejnih sistemov.

V zadnjem času se poleg govornih signalov, njihovih prepisov in slovarja izgovarjav govornim zbirkam dodajajo tudi drugi opisi govorcev, kot so npr. paralingvistična stanja. Eden takih dodatnih opisov so tudi čustvena stanja govorca.

Čustvene govorne podatkovne zbirke, ki so razvite za uporabo razvoja samodejnih sistemov za tvorjenje govora ali razpoznavanje čustvenega stanja govorca, lahko zgradimo s pomočjo dveh pristopov. Prvi je snemanje govorne zbirke s poklicnimi govorci, ki so zmožni igrati emocionalna stanja. Take zbirke so posnete z vnaprej pripravljenimi povedmi, ki so izbrane iz obsežnejših zbirk besedil in

skušajo zadostiti fonetični porazdelitvi osnovnih enot posameznega jezika. V drugem primeru pa je razvoj zbirke osredotočen na pridobivanje že posnetih govornih segmentov. Pomemben dejavnik, ki dodatno označuje govorno gradivo, zajeto v zbirko govora, je spontanost. Tako lahko zasledimo zbirke, ki vsebujejo predvsem posnetke spontanega govora, in zbirke govora, ki zajemajo igran oz. bran govor. Če si želimo pridobiti govorno zbirko čustvenega govora, moramo posnetkom pripisati tudi čustveno kategorijo oz. čustveno stanje govorca, ki ga odraža govor v posameznem posnetku. Postopek pripisa čustvene oznake se razlikuje glede na način zajema govornega gradiva. Pri prvem, ki predvideva novo snemanje govorne zbirke s poklicnimi govorci, so čustvene oznake posameznega posnetka določene vnaprej. Drug način, ki predvideva razčlembo in prepis že posnetih govornih signalov, pa zahteva poznejše označevanje čustvenih stanja govorca. V zadnjem času se za to nalogo velikokrat najamejo označevalci, katerih večinsko mnenje določa končno oznako posameznega posnetka. Označevalci lahko pomagajo tudi pri vrednotenju že določenih oznak. Na podlagi večinskega mnenja lahko razvijalci vrednotijo uspešnost posnemanja čustvenega stanja tudi pri posnetkih poklicnega govorca.

1.1 Pregled obstoječih zbirk čustvenega govora

Do danes je bilo razvitih veliko tujejezičnih govornih zbirk, ki skušajo zajeti tudi paralingvistična stanja govorca (Schuller idr. 2013). Tovrstna stanja se v literaturi opisujejo kot stanja govorca, ki se ne dajo opisati z lingvističnimi ali fonetičnimi oznakami. Lahko so izražena v govoru, kot na primer omotičnost, razpoloženje, zanimanje, čustveno stanje itd. Načrtovanje graditve tovrstnih podatkovnih zbirk zahteva kompleksno interdisciplinarno sodelovanje. Eden pomembnejših dejavnikov je prav opredelitev paralingvističnih oznak, kjer je nujno potreben strokovnjak predvidenega področja uporabe podatkovne zbirke. Označevanje čustvenih stanj v govoru je težavna naloga, saj trenutno nimamo splošno uveljavljene metodologije opisovanja čustvenih stanj. V takem primeru se raziskovalci velikokrat zatečejo k utečenim postopkom graditve govornih podatkovnih zbirk po zgledih v svetovni literaturi, ki natančno opredeljujejo opise čustve-

nih stanj v govoru in se glede na potrebe raziskovanja močno razlikujejo. V literaturi (Cowie in Cornelius 2003) zasledimo osnovne raziskovalne smernice za graditev govornih zbirk čustvenega govora, ki so osredotočene predvsem na določen raziskovalni cilj. Največkrat se takšne zbirke pridobivajo za raziskovanje teoretskega ozadja čustvenih stanj v govoru, ki so v večini primerov psihološke ali biološke narave. Na drugi strani pa lahko takšne zbirke pridobivamo tudi za razvoj različnih aplikacij govornih tehnologij.

Za slovenski jezik obstajata dve zbirki čustvenega govora (Gajšek idr. 2009a; Hozjan idr. 2002). Prva je večmodalna zbirka spontanih čustvenih stanj, druga pa je del večjezične zbirke igranega govora, Interface. Ta je dostopna pod komercialno licenco. Žal ni vsaka govorna zbirka primerna za uporabo v sintezi govora. Eden odločilnih parametrov je količina govornih posnetkov posameznega govorca. Za sintezo je potrebno, da je posnetkov enega govorca čim več, pri čemer naj bi govorec zajel čim večji besedni zaklad jezika, v katerem govori. Obe omenjeni govorni zbirki zato za sintezo nista primerni.

2 METOLOGIJA GRADITVE ZBIRKE EMOLUKS

Zbirka čustvenega slovenskega govora je bila razvita v Laboratoriju za umetno zaznavanje, sisteme in kibernetiko (LUKS) pod imenom EmoLUKS. Njen razvoj temelji na dolgoletnih izkušnjah pri razvoju slovenskih govornih podatkovnih zbirk (Mihelič idr. 2003). Zbirka je zasnovana na posnetkih slovenskih radijskih iger, ki smo jih pridobili od RTV Slovenija za akademsko uporabo. Vsebuje natančne prepise govornih posnetkov hkrati z oznakami čustvenih stanj govorcev.

2.1 Razčlemba in prepis radijskih iger

S pomočjo RTV Slovenija smo pridobili zvočne posnetke in scenarije 17 radijskih iger, ki so bili v večinoma narejeni v profesionalnem studiu Radia Slovenija. Vsako igro smo transkribirali ter razčlenili glede na identiteto govorca. V veliko pomoč so nam bili pri prepisu govornega gradiva scenariji iger.

Za potrebe prepisov in razčlenitve glede na govorca smo uporabili program

Transcriber (Barras idr. 2001).

Program omogoča hitro in učinkovito razčlenjevanje govornih signalov glede na govorce, njihovo transkribcijo in označevanje nejezikovnih delov govora v posnetku. Posnetke smo razčlenili tudi glede na zaključene stavčne enote. S takim pristopom smo pridobili nabor posnetkov, ki niso predolgi in hkrati dajejo dovolj konteksta za označevanje paralingvistične informacije v govoru.

Označili smo 17 posnetkov radijskih iger v približnem skupnem časovnem obsegu 12 ur in 50 minut. Tabela 1 prikazuje količino transkribiranega in označenega gradiva.

Št.	**Naslov radijske igre**	**Trajanje**
1	Penzion Evropa	0:48:03,56
2	Angleško poletje	0:57:55,69
3	V Sieni nekega deževnega dne	0:42:32,59
4	Aut Caesar	0:33:22,25
5	Štefka	0:36:45,69
6	Podzemne Jame	0:46:17,56
7	Na glavi svet	0:58:29,32
8	Naš novi najboljši prijatelj	0:26:51,25
9	Dediščina	0:54:36,62
10	Potovalci	0:49:50,27
11	Nič brez Deteljnika	0:48:00,00
12	Sokratov zagovor	1:09:51,34
13	Nedotakljivi – Četrti žebelj	0:38:44,35
14	Nedotakljivi – Moj ded Jorga Mirga	0:37:00,00
15	Nedotakljivi – Moj oče Ujaš Mirga	0:40:04,45
16	Nedotakljivi – Jaz, Lutvi Belmondo aus Shangkai Gav	0:35:09,83
17	Hipopituitarizem ali namišljeni bolnik	0:46:50,35
	Skupaj	12:50:25,12

Tabela 1: Pregled trajanja 17 radijskih iger.

Razvijalci samodejnih sistemov za razpoznavanje ali tvorjenje umetnega govora želijo razpolagati z govornimi podatkovnimi zbirkami, ki vsebujejo predvsem

čist govor. Zato smo zbirko zasnovali tako, da smo med razčlenjevanjem in zapisovanjem besed vzporedno označevali tudi nejezikovne prvine, ki so večkrat del radijskih iger, kot npr. glasba v ozadju, različni šumi in raznovrstni dodatni zvočni učinki. Poleg tega nismo pozabili na druge nejezikovne prvine govorca, kot so vdih, cmokanje, stokanje, jok in smeh. Govorno gradivo, namenjeno označevanju in nadaljnji obdelavi, je v povprečju za polovico časa krajše, kot ga izgovori posamezni igralec v radijski igri.

Transkribirano in razčlenjeno gradivo obsega 45 moških govorcev in 23 ženskih govork ter en otroški glas. Kvantitativna predstavitev, ki je podrobneje predstavljena v Justin (2016) ponazarja, da lahko iz vsega označenega gradiva v radijskih igrah uporabimo le 46 odstotkov govora. Gre za čist govor, ki ne vsebuje drugih motenj v posnetku, kot so glasba v ozadju, drugi slušni efekti ali pa hkratni govor večjega števila igralcev.

2.2 Izbira čustvenih stanj za označevanje v zbirki EmoLUKS

Označevanje čustvenih stanj v govoru poteka s pomočjo izvedenskega znanja. Govornim posnetkom lahko pripiše oznako izvedenec za dano področje. V zadnjem času pa se čedalje pogosteje uporablja nabor označevalcev, ki podajo mnenje o posameznem posnetku. S takim naborom mnenj lahko bolj splošno določimo oznako posnetku. Ker je označevanje govornih posnetkov velikokrat dolgotrajen proces, čedalje pogosteje uporabljamo spletne aplikacije, ki omogočajo podajanje mnenj in/ali označevanje govornih ali video posnetkov. Tak pristop zagotavlja hkratno označevanje večjega števila označevalcev in obenem ponuja označevalcem svobodno izbiro časovnega okvira označevanja. V literaturi zasledimo tak pristop pod pojmom množično izvajanje (ang. crowd-sourcing) (Howe 2006).

Avdio- ali videogradivo je vedno določeno s časovno komponento, zato lahko ponudimo označevalcem, da sami izberejo segment označenega gradiva, ali pa sami pripravimo gradivo v smiselnih odsekih. Pri označevanju čustvenih stanj v govoru (Gajšek idr. 2009a; Douglas-Cowie idr. 2003) zasledimo obe izvedbi. Prva poudarja označevanje vnaprej pripravljenih odsekov posnetkov, druga pa

prepušča določanje časovne razmejitve označevalcem glede na njihovo prepričanje o začetku in koncu določenega čustvenega stanja govorca. Pri večji količini govora različnih govorcev, predvsem pa tedaj, ko se čustvena stanja govorca pojavljajo v dialogu med dvema ali večjim številom govorcev, morajo označevalci nameniti dodatno pozornost razčlembi govora tudi med različne identitete govorcev. Po našem prepričanju je to za prostovoljne označevalce prezahtevna naloga, saj zahteva za označevanje veliko večjo zbranost in posledično več časa. Tovrstni pristop morda omogoča večjo zanesljivost in ujemanje označevalcev na krajših odsekih govornega signala, hkrati pa od označevalca zahteva veliko spretnosti in dobro poznavanje aplikacije za tovrstno razmejevanje in hkratno označevanje čustev v posnetkih. Glede na pridobljeno govorno gradivo, ki je zastopano v radijskih igrah, smo se odločili, da označevalcem ponudimo že razmejene smiselne odseke govornega gradiva, katerim le pripišejo svoje mnenje o čustvenem stanju govorca. Težava tovrstnega pristopa se pokaže pri označevanju krajših povedi. V takih primerih se označevalec težko ali celo ne more odločiti, s katero čustveno kategorijo bi lahko opisal posnetek. Da bi se temu izognili, lahko označevalcu ponudimo tudi branje širšega konteksta povedi, ki jo označuje. Zato smo za boljše razumevanje konteksta posnetka, v katerem govorec izraža čustveno stanje, predvideli zapis nekaj predhodnih in sledečih povedi.

Po pregledu literature in preizkusu dostopnih spletnih aplikacij, ki omogočajo označevanje govornih posnetkov, smo ugotovili, da nobena v taki meri ne izpolnjuje pogojev, ki bi morali biti upoštevani, da bi lahko prostovoljnim označevalcem omogočili kakovostno in hitro označevanje. Zato smo se odločili izdelati spletno aplikacijo, namenjeno označevanju zvočnih ali videoposnetkov. K taki odločitvi nas je napeljalo tudi dejstvo, da smemo podatke uporabljati samo za akademske potrebe. Podrobnejša predstavitev aplikacije je podana v Justin (2016).

Vsako raziskovalno delo, ki posega na področje čustvenih stanj, potrebuje najprej definicijo, ki opisuje posamezno čustveno stanje, in skuša opredeliti, kaj bo glavni vzvod s širokega področja definicij in nazorov za proučevanje čustve-

nega stanja človeka. Razlike v teoretičnih ozadjih, ki opisujejo čustvena stanja, pričajo o različnem pojmovanju čustvenih stanj (Cornelius 1996). Lahko jih delimo na štiri različne poglede (Cornelius 2000). Uporaba vsakega od njih narekuje tudi različno pojmovanje povezav med definiranimi čustvenimi kategorijami. Ko želimo čustvena stanja označevati, jih obdelovati ali razvrščati v skupine, moramo upoštevati predvsem značilnost, ki omogoča umestitev posameznih čustvenih kategorij v skupine. Posamezna čustvena kategorija pomeni čustveno stanje, katere člani so si med seboj podobni bolj kot s člani drugih kategorij. Različni pogledi na modeliranje posameznih relacij med emocionalnimi zvezami so podrobneje predstavljeni v Justin (2016) in Cornelius (1996).

V prispevku se osredotočamo na razdelitev čustvenih stanj po Darwinovem pogledu. Predpostavljamo, da obstajajo osnovna čustvena stanja, ki jih lahko predstavimo v diskretnem modelu čustvenih kategorij. Čustvene kategorije, ki smo jih označevali v zbirki EmoLUKS, so: žalost, veselje, gnus, jeza, strah in presenečenje. Tak pristop je eden pogosteje uporabljenih za modeliranje čustvenih stanj govorca. Tem kategorijam smo dodali tudi nevtralno čustveno stanje in oznako »nič-od-tega«, ki pomeni stanje govorca, ki ga označevalec ne more natančno določiti, saj se govorec po oceni ocenjevalca nahaja v eni izmed zahtevnejših kategorizacij čustvenih stanj, kot je na primer Plutchnikov diskretni model čustev (Plutchnik 1962).

2.3 Postopek označevanja zbirke čustvenega govora EmoLUKS

V prejšnjih razdelkih smo opisali korake, ki so pri tovrstni zasnovi zbirke čustvenega govora pogoj, da lahko ponudimo posamezne posnetke označevalcem v označevanje. Ker nameravamo zbirko uporabiti predvsem pri proučevanju umetnega tvorjenja čustvenega govora, smo označevalcem čustev v govoru ponudili le posnetke ene govorke in enega govorca. Na to so nas napeljala dejstva, ki jih narekuje časovna analiza razčlenjenega in transkribiranega govora. Za potrebe umetnega tvorjenja govora si na splošno želimo razpolagati s čim več posnetega govora enega govorca in s čim večjo zastopanostjo posnetkov čim bolj raznolikega govora v vseh čustvenih kategorijah. Kvantitativni pregled raz-

členjenega in zapisanega govornega gradiva radijskih iger nam narekuje izbiro govorca z oznako 01m_av in govorke z oznako 01f_lb.

Označevanje podatkovne zbirke EmoLUKS je potekalo v dveh ločenih iteracijah. Za tak postopek smo se odločili na podlagi rezultatov, ki smo jih pridobili v prvem koraku označevanja in smo o njih poročali v Justin idr. (2014). V prispevku smo navedli težave pri označevanju čustvenih stanj v govoru in poročali o rezultatih, pridobljenih s petimi označevalci. Iz prispevka je razvidno, da do popolnega konsenza med označevalci prihaja le v redkih primerih. V označenem gradivu ga zasledimo med petimi in desetimi odstotki v posamezni čustveni kategoriji. Čeprav je odstotek popolnega konsenza med označevalci majhen, lahko vseeno potrdimo, da igralci v radijskih igrah jasno izražajo čustvena stanja in da je izbira takih posnetkov govora smiselna za graditev slovenske zbirke čustvenega govora. Vseeno pa moramo poudariti, da sta priprava in poznejše označevanje čustvenih stanj dolgotrajen proces, ki žal ne more zagotoviti strinjanja označevalcev v tolikšni meri, kot to lahko zasledimo pri zbirkah čustvenega govora, ki so bile zajete s pomočjo namenskega snemanja čustveno obarvanih in vnaprej pripravljenih povedi (Hozjan idr. 2002).

Z analizo prve iteracije označevanja smo ugotovili, da je pri 17 odstotkih vseh označenih posnetkov čustveno stanje govorca glede na večinsko mnenje označevalcev nedoločeno. Od vseh nedoločenih posnetkov je 91 odstotkov primerov takih, ko sta jim dva označevalca pripisala eno čustveno stanje, druga dva drugo, peti pa tretje. Preostalih 9 odstotkov v naboru vseh čustveno nedoločenih posnetkov pa je takih, da jim je vsak od petih označevalcev pripisal drugačno čustveno stanje. Zato smo se odločili, da za te problematične posnetke ponovimo označevanje. V ta namen smo k označevanju znova povabili istih pet označevalcev. Ponovno označevanje je potekalo v drugem časovnem obdobju in ga v nadaljevanju imenujemo druga iteracija označevanja. S takim pristopom smo hoteli zmanjšati količino nedoločenih posnetkov ter preveriti konsistentnost odločitev označevalcev. Tako smo želeli tudi ugotoviti, ali označeni posnetki resnično vsebujejo večdimenzionalna oz. prepletajoča se čustvena stanja govorcev v radijskih igrah.

Druga iteracija je vsebovala posnetke iste govorke in govorca kot prva. Vanjo smo zajeli predvsem posnetke, ki jim ni bilo mogoče pripisati čustvene oznake na podlagi večinske odločitve označevalcev. Poleg takih smo v drugo iteracijo označevanja vključili tudi posnetke, ki so bili v neposredni bližini posnetkov, ki jim stanja ni bilo mogoče določiti. S takim pristopom smo zagotovili tudi ponovno obravnavo posnetkov, za katere menimo, da so težje določljivi, saj se v radijskih igrah, kjer je prisoten predvsem dialog med igranimi osebami, čustvena stanja izrazijo tudi prek daljšega odseka besedila. V prvi iteraciji so lahko označevalci poročali o napakah, ki so bile del zapisa ali pa nenatančne razčlenitve govornega signala. Ne glede na označbo večinskega mnenja označevalcev smo vključili v drugo iteracijo tudi vse posnetke, kjer smo napake v transkribciji in razčlenitvi odpravili. Po pregledu rezultatov prve iteracije in izračunu trajanja vsakega posnetka, ki je bil del prve iteracije, smo tudi ugotovili, da ločnica (ločilna meja med posnetki), ki je pomenila zaključek govorne enote, ni povsem primerna za označevanje čustvenih stanj. V prvi iteraciji smo za ločnico govornega signala uporabljali celo poved. Pri daljših povedih pa se v radijskih igrah izkaže, da igralci lahko izražajo tudi več čustvenih stanj. Daljše povedi, ki vsebujejo večdimenzionalna stanja ali pa se lahko čustva v njih tudi povezujejo, niso primerne za predlagani način označevanja. Enega izmed vzrokov za velik delež nedoločenih čustvenih stanj govorca, ki so jih označevalci označili v prvi iteraciji, lahko pripišemo tudi predolgim posnetkom oziroma ločilni meji, ki smo jo izbrali za razčlembo posnetkov. Zaradi preverjanja te hipoteze smo določene dolge povedi smiselno razčlenili na več odsekov.

V obeh iteracijah označevanja je sodelovalo pet označevalcev, od tega trije moški in dve ženski. Vsak je pri podajanju oznak uporabljal slušalke in si sam izbiral, kdaj je posnetke označeval.

2.4 Kvantitativni pregled posnetkov v postopku označevanja

Pripisane čustvene oznake posnetkom v zbirki EmoLUKS izhajajo iz dveh iteracij označevanja, zato je ključnega pomena jasno in natančno opredeliti zastopanost in trajanje posnetkov v vsaki iteraciji označevanja ter predstaviti način

združevanja medsebojno neodvisnih posnetkov v zbriko EmoLUKS.

Posnetke, ki so bili zajeti v drugo iteracijo označevanj, lahko razdelimo v tri skupine. Prva so posnetki, katerih ločilna meja ni bila spremenjena in pomeni celotno poved. Druga so posnetki, ki smo jih v postopku razločevanja pridobili s podrobnejšo razčlenitvijo nekaterih povedi. Tretja skupina so posnetki, ki so bili označeni v prvi ali drugi iteraciji in so med seboj popolnoma neodvisni.

Skupina	**Govorec**	**1. iteracija**		**2. iteracija**		**EmoLUKS**	
		Št. pos.	**T [s]**	**Št. pos.**	**T [s]**	**Št. pos.**	**T [s]**
Polna povezanost	01m_av	387	1377	387	1364	387	1364
	01f_lb	155	328	155	315	155	315
	skupaj	542	1705	542	1679	542	1679
Razdeljena povezanost	01m_av	133	1512	327	1505	327	1505
	01f_lb	38	190	94	183	94	183
	skupaj	171	1702	421	1688	421	1688
Brez povezave	01m_av	242	799	19	83	261	883
	01f_lb	155	377	6	9	161	387
	skupaj	397	1177	25	92	422	1270
Skupaj	01m_av	762	3689	733	2952	975	3752
	01f_lb	348	896	255	507	410	885
	skupaj	1110	4584	988	3459	1385	4636

Tabela 2: Primerjava količine posnetkov za prvo in drugo iteracijo označevanja in količina končne zbrane zbirke EmoLUKS. V tabeli je trajanje označeno s črko *T* in je izraženo v sekundah.

V Tabeli 2 povzemamo število posnetkov in njihovo skupno trajanje za vsako iteracijo posebej. Smiselno združimo posnetke v podatkovno zbirko EmoLUKS, ki je kvantitavno opredeljena v zadnjem stolpcu v Tabeli 2. Zaradi nazornega pregleda nad združevanjem posnetkov iz prve in druge iteracije označevanja delimo količino posnetkov na podsklope, katerih izvor se kaže v prej omenjenih skupinah, ki jih imenujemo »polna povezanost«, »razdeljena povezanost« in »brez povezave«.

Oznake v zbirki EmoLUKS sestavljajo med seboj neodvisne oznake posnetkov,

ki so bili označeni v prvi ali drugi iteraciji označevanja. Tako zbirka EmoLUKS vsebuje 1385 posnetkov čistega govora ene govorke in enega govorca v skupnem času 1 ure 17 minut in 16 sekund. Povprečni čas trajanja posnetka je 3,3 sekunde, mediana trajanja pa 2,3 sekunde.

3 ANALIZA OZNAČENEGA GOVORNEGA GRADIVA

3.1 Ujemanja oznak označevalcev

Pri graditvi podatkovnih zbirk, ki vsebujejo čustvene oznake na podlagi govornih in/ali videoposnetkov, ponavadi nimamo na voljo referenčne čustvene oznake, ki bi omogočala preverjanje oznak označevalcev. Zato samo analizo in hkrati končni pripis čustvene oznake posameznemu posnetku lahko predstavimo z ujemanjem oznak večjega števila označevalcev. Na splošno si želimo pridobiti čim več posnetkov, kjer bi se ocenjevalci povsem strinjali o čustveni oznaki posameznega posnetka. Ravno pri označevanju čustev je to zahtevna naloga, saj trenutno nimamo splošno uveljavljene definicije, kaj točno čustveno stanje je, in so zato posameznikove oznake čustev odvisne od subjektivnega mnenja posameznika. To ponavadi privede do popolnega ujemanja oznak v manj primerih kot v primerjavi z drugo skrajnostjo, kjer imamo popolno nestrinjanje označevalcev. Za analizo ujemanja mnenj označevalcev se običajno uporablja κ statistika.

Čustvene oznake ocenjevalcev so nominalni kvalitativni podatki. Kriterij kvalitativnosti lahko predstavimo kot kriterij, ki opredeljuje podatke, nad katerimi ne moremo izvajati računskih operacij. Kriterij nominalnosti pa izhaja iz medsebojne primerjave kategorij. Pri čustvenih oznakah ne moremo reči, katera od oznak (kategorij) je večja ali manjša, in jih posledično ne moremo urejati, zato tovrstne oznake lahko opredelimo kot nominalne.

V zadnjih petdesetih letih so bile predstavljene različne vrste merjenja κ koeficientov za nominalne kvalitativne podatke (Cohen 1960; Fleiss 1971; Randolph 2005). Vsaka ima svoje omejitve uporabe. Vse pa izhajajo iz osnovne enačbe (1) za izračun κ koeficienta po Cohenu (Fleiss 1971). Cohen je leta 1960 opredelil

izračun κ koeficienta le za dva označevalca, pri čemer P_e označuje pričakovano naključno ujemanje med označevalcema (ang. agreement between raters expacted by chance), P_O pa dejansko razmerje ujemanja (ang. overall observed aggrement). Vrednost κ koeficienta je predstavljena kot razmerje med dejanskim izmerjenim ujemanjem med označevalcema in ujemanjem, ki bi ga dosegla dva označevalca, če bi označevala naključno.

$$\kappa = \frac{P_O - P_e}{1 - P_e} \tag{1}$$

Zaloga vrednosti κ koeficienta je omejena na interval med -1 in 1. Vrednost 0 pomeni naključno ujemanje, vrednost 1 pomeni popolno ujemanje, negativne vrednosti pa pomenijo odstopanje od pričakovanega. Pozitivne vrednosti κ koeficientov sta kategorizirala Landis in Koch (1977) in še zdaj veljajo za splošno sprejeto kategorizacijo brez kakršnega koli dokaza. Landis in Koch sta glede na vrednost κ koeficientov podala opisno kategorizacijo. Negativne κ vrednosti so predstavljene kot slabo ujemanje, $0,01 \leq \kappa \geq 0,20$ kot rahlo ujemanje, $0,21 \leq \kappa \geq 0,40$ kot pošteno ujemanje, $0,41 \leq \kappa \geq 0,60$ kot zmerno ujemanje, $0,61 \leq \kappa \geq 0.80$ kot znatno ujemanje in $0,81 \leq \kappa \geq 1,00$ kot skoraj popolno ujemanje. Ker kategorizacija ni podprta z dokazom, jo raziskovalci navajajo le kot oporo za lažje opredeljevanje ujemanja in je uporabljena tudi za druge posplošitve Cohenove enačbe za izračun κ koeficienta.

Definicija za izračun κ koeficienta po Cohenu ni primerna za izračun ujemanja med večjim številom ocenjevalcev. Posplošitev Cohenove definicije je leta 1971 predstavil Fleiss (1971). Posplošitev izhaja iz definicije parametra P_O in P_e v enačbi (1), zgoraj.

Fliess je parameter P_O definiral, kot ga podajamo v enačbi (2).

$$P_O = \frac{1}{Nn(n-1)} \left(\left(\sum_{i=1}^{N} \sum_{j=1}^{k} {n_{ij}}^2 \right) - Nn \right), \tag{2}$$

kjer je n_{ij} število označevalcev, ki je vzorec i pripisalo razredu j, n število vseh

označevalcev, N število vseh vzorcev in k število vseh kategorij (čustvenih oznak). Parameter P_e je definiran v enačbi (3), kjer so pomensko označeni enaki simboli kot v enačbi (2).

$$P_e = \sum_{j=1}^{k} \left(\frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \right)^2 \tag{3}$$

Fliessov izračun κ koeficienta podaja zanesljivost ujemanja med določenim številom ocenjevalcev z uporabo števila ocen v razmerju s kategorijami razpoznavanja. Za izračun Fliessove κ statistike moramo imeti na voljo binarno ali nominalno porazdelitev vzorcev. Pomanjkljivost za uporabo natančnega izračuna ujemanja mnenj označevalcev pri označevanju čustvenih stanj v govoru z uporabo Fliessove κ statistike najdemo v predpostavki, da morajo biti podatki v vseh kategorijah k enakomerno zastopani. Če predpostavka ni izpolnjena, vrednost κ po Fliessu drastično pade, in sicer neodvisno od števila primerov, kjer so se označevalci popolnoma strinjali. Pri označevanju čustvenih stanj v govoru imamo ponavadi nesimetrično zastopanost čustvenih kategorij, saj večkrat prevladuje nevtralni govor, zato tudi Fliessova definicija izračuna κ koeficienta za podajanje ujemanja mnenj označevalcev v zbirki EmoLUKS ni primerna.

Rešitev najdemo v delu Randolph (2005), ki predlaga modifikacijo izračuna Fliessovega κ koeficienta. Razlika je v definiciji parametra P_e, ki ga nadomesti z enačbo (4), kjer k pomeni število kategorij (tj. število čustvenih oznak).

$$P_e = \frac{1}{k} \tag{4}$$

S tako definiranim parametrom P_e se izognemo apriorni omejitvi glede enakomerne porazdelitve razredov in s tem posledično težavam zaradi kvadratnega vpliva prevladujočega razreda oznak na izračunano vrednost κ koeficienta. Izračun κ koeficienta po Randolphu, ki ga v svojem delu imenuje mnogooznačevalski prostoprostostni kappa koeficient κ_{free} (ang. multirater free-marginal kappa), je v celoti definiran z enačbo (5).

$$\kappa_{free} = \frac{\left(\frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2 - Nn\right)\right) - \frac{1}{k}}{1 - \frac{1}{k}} \tag{5}$$

Randolphov izračun κ koeficienta pa lahko v določenem primeru pomeni tudi slabost. Število vseh mogočih kategorij, ki jih uporabljamo pri označevanju, postane pomembno in vpliva na končno vrednost izračunanega κ koeficienta. Vrednost koeficienta $\frac{1}{k}$ pada z večanjem števila mogočih kategorij, kar pa vpliva na večje vrednosti κ koeficienta tudi pri stalnih vrednostih podatkov (n_{ij} , n, N). Zato je pred uporabo izračuna κ koeficienta potreben dober premislek, katere kategorije ponuditi označevalcem v označevanje.

Vsi opisani postopki za izračun κ koeficientov obravnavajo vsak vzorec v podatkih z enako utežjo, kar se v končnem rezultatu odraža z dejstvom, da vsak vzorec prispeva enako utežen prirastek. Pri zbirkah, ki zajemajo posnetke govornega ali videogradiva, pa je vsak vzorec določen z začetnim in končnim časom. Z izračunom zgoraj opisanih κ koeficientov na primeru zbirke čustvenih stanj govorca končni rezultat upošteva le število vseh posnetkov, ki so ocenjeni z označevalci, ne pa tudi njihovega trajanja. Izražanje čustev je odvisno tudi od trajanja čustvenega stanja, v katerem se govorec nahaja. Zato menimo, da je smiselno k prirastku posameznega posnetka pri vrednosti izračunanega κ koeficienta upoštevati tudi njegovo trajanje. V ta namen so Gajšek idr. (2009b) predlagali časovno utežen κ koeficient (ang. time-weighted kappa coefficient, κ_{tw}). Prednosti njegove uporabe so predstavili na primeru zbirke spontanega čustvenega govora AvID (Gajšek idr. 2009a). Poudariti je treba, da so pri označevanju čustvenih stanj imeli označevalci prosto izbiro časovnega intervala za označevanje čustvenega stanja govorca, zato se je tak izračun κ koeficienta izkazal za še posebej uspešnega. Za razlago izračuna časovno oteženega κ koeficienta se lahko spet ozremo na enačbo (1), ki jo je definiral Cohen. Parameter P_e ohranja enak izraz kot pri definiranju po Randalphu v enačbi (4). S tako izbiro se izognemo opisani problematiki pri nesimetrično zastopanih kategorijah. Parameter P_O pa namesto enakomernega povprečenja vseh vzorcev izraža pov-

prečje glede na dolžino posameznega vzorca. Tako je definiran parameter P_O z enačbo (6), pri čemer T pomeni celotno trajanje vseh vzorcev, t_i pa trajanje posameznega vzorca.

$$P_O = \frac{1}{Tn(n-1)} \sum_{i=1}^{N} \left(\sum_{j=1}^{k} n_{ij}^2 - n \right) t_i \tag{6}$$

Končni izraz κ_{tw} koeficienta pa je predstavljen z enačbo (7).

$$\kappa_{tw} = \frac{\left(\frac{1}{Tn(n-1)} \sum_{i=1}^{N} \left(\sum_{j=1}^{k} n_{ij}^2 - n \right) t_i \right) - \frac{1}{k}}{1 - \frac{1}{k}} \tag{7}$$

Govorno gradivo, ki je bilo označeno v dveh različnih iteracijah označevanja, v nadaljevanju analiziramo in predstavimo ujemanje mnenj označevalcev. Kot je opisano v zgornjih odstavkih, so zaradi nesimetrične zastopanosti označenih čustvenih kategorij smiselni le izračuni κ koeficientov po Randolphovi definiciji v enačbi (5) in tudi s časovno uteženim κ koeficientom, ki je definiran z enačbo (7). S tako pridobljeno analizo želimo opozoriti na različno pojmovanje čustvenih stanj pri posameznem označevalcu ter hkrati opozoriti na različne odločitve označevalcev pri istih posnetkih v dveh različnih časovnih obdobjih.

V Tabeli 3 lahko medsebojno primerjamo le skupino z imenom »polna povezanost«, in to le med prvo in drugo iteracijo označevanja. V tej skupini je bilo označenih 543 posnetkov, od tega 378 posnetkov govorca 01m_av in 155 posnetkov govorke 01f_lb. Označevanje je izvedlo istih pet označevalcev. Pri ponovnem označevanju (2. iteracija) opazimo večje ujemanje med ocenjevalci, še zdaleč pa ne tako znatnega, da bi lahko drugače kategorizirali ujemanje, kot je to predstavljeno v Landis in Koch (1977). Glede na pridobljene κ vrednosti v Tabeli 3 lahko za vsako skupino in tudi za vse združene podatke kategoriziramo ujemanje med označevalci kot »pošteno« ujemanje.

Skupina	Govorec	1. iteracija		2. iteracija		EmoLUKS	
		$\kappa_{\mathbf{free}}$	$\kappa_{\mathbf{tw}}$	$\kappa_{\mathbf{free}}$	$\kappa_{\mathbf{tw}}$	$\kappa_{\mathbf{free}}$	$\kappa_{\mathbf{tw}}$
Polna	01m_av	0,31	0,32	0,32	0,36	0,33	0,35
pove-	01f_lb	0,29	0,28	0,35	0,35	0,34	0,33
zanost	skupaj	0,30	0,31	0,33	0,36	0,33	0,35
Razdelj-	01m_av	0,31	0,35	0,31	0,36	0,30	0,35
ena pove-	01f_lb	0,30	0,34	0,29	0,30	0,30	0,31
zanost	skupaj	0,30	0,35	0,31	0,35	0,30	0,34
Brez	01m_av	0,40	0,42	0,34	0,34	0,46 (0,39)	0,48 (0,41)
povezave	01f_lb	0,38	0,37	0,22	0,24	0,44 (0,37)	0,44 (0,37)
	skupaj	0,39	0,41	0,31	0,33	0,45 (0,39)	0,47 (0,40)
Skupaj	01m_av	0,34	0,36	0,32	0,36	0,35	0,38
	01f_lb	0,33	0,33	0,33	0,33	0,37	0,38
	skupaj	0,34	0,35	0,32	0,35	0,36	0,38

Tabela 3: Primerjava ujemanja označevalcev s koeficientom κ_{free}, in koeficientom κ_{tw}. V skupini »brez povezave« so v oklepajih predstavljeni izračuni κ koeficientov z realnimi petimi označevalci, medtem ko vrednosti, ki niso v oklepajih, pomenijo vrednosti nad združenimi posnetki, katerih mnenja ocenjevalcev v posamezni kategoriji smo upoštevali dvakratno.

3.2 Konsistentnost mnenj označevalcev

Primerjavo podanih mnenj posameznega ocenjevalca v prvi in drugi iteraciji lahko opazujemo v tabeli zamenjav (ang. confussion matrix) za posameznega označevalca. Obsežne tabele zamenjav so predstavljene v Justin (2016), v tem prispevku predstavljamo rezultate strnjeno v eni sami Tabeli 4. Predstavljeni rezultati so zbrani le za skupno zastopanost označenih posnetkov, ki so jim ocenjevalci mnenje v drugi iteraciji spremenili, in hkrati skupno zastopanost označenih posnetkov, ki jim ocenjevalci niso spremenili čustvenih oznak. Dodatno v tabeli zaradi boljše nazornosti izračunamo tudi delež spremenjenih in nespremenjenih mnenj označevalcev.

Rezultate zopet razdelimo glede na izvor posnetkov. Tokrat lahko opazujemo spremembe le za skupino »polno povezanih« in skupino »razdeljena povezanost«. Druga skupina v tabelah zamenjav in Tabeli 4 so rezultati na podlagi

naslednjega predvidevanja: če daljše posnetke, v katerih je jasno izraženo določeno čustveno stanje govorca, razdelimo na manjše odseke, bodo zaradi jasnosti izražanja čustvenega stanja označevalci pripisali enako čustveno stanje v vseh ponovno označenih razdelnih posnetkih. Tako skupina »razdeljena povezanost« pomeni rezultate vsakega razdeljenega posnetka, ki je bil označen v drugi iteraciji, pri čemer kot referenčno oznako predvideva mnenje posameznega označevalca iz prve iteracije.

Tabela 4 priča o kompleksnosti označevanja čustvenih stanj govorca. Povprečno število vseh spremenjenih mnenj označevalcev v prvem sklopu posnetkov pokaže, da so označevalci v povprečju kar v 49 odstotkih primerov spremenili svoje mnenje glede na podano mnenje v prvi iteraciji. Če opazujemo spremembo mnenj pri posameznem ocenjevalcu neodvisno od identitete govorca (razdelek skupaj), opazimo, da sta ženski označevalki v povprečju zamenjali mnenje kar v 57 odstotkih, moški označevalci pa v 43 odstotkih primerov. Na podlagi rezultatov v obravnavanem podsklopu posnetkov lahko rečemo, da je konsistentnost petih označevalcev, ki so označevali čustvena stanja govorcev iz posnetkov radijskih iger, slaba.

Pri drugi skupini posnetkov lahko preverimo, ali daljši posnetki, katerih dolžino določa celotna poved, jasno izražajo čustvena stanja govorca in, ali se čustvena stanja govorca spreminjajo tudi v krajših smiselnih stavkih. V Tabeli 4 opazimo, da je povprečna sprememba mnenja pri vseh označevalcih zaznana v 56 odstotkih primerov. Zopet sta ženski ocenjevalki večkrat spremenili mnenje kot moški ocenjevalci. Povprečen delež sprememb vseh zajetih posnetkov v podsklopu »razdeljena povezanost« kaže, da so v daljših povedih pri radijskih igrah lahko izražena tudi čustvena stanja govorca, ki se medsebojno prepletajo ali celo spreminjajo. Na podlagi tega lahko sklepamo, da so za označevanje čustvenih stanj primernejše krajše smiselne zaključene enote.

Tabela 4 podaja tudi skupno število sprememb mnenj ocenjevalcev. V povprečju so izmed 963 posnetkov, ki so bili v celoti ocenjeni tudi v prvi iteraciji označevanja, spremenili svojo odločitev kar v 52 odstotkih primerov. To dejstvo naka-

zuje na težavno dojemanje čustvenih stanj govorcev v radijskih igrah in seveda dodatno potrjuje vzrok, zaradi katerega ni splošno sprejete definicije čustvenih stanj pri človeku.

Označevalec	Govorec	Polna povezanost				Razdeljena povezanost				Skupaj			
		št. nespr.	[%]	št. spr.	[%]	št. nespr.	[%]	št. spr.	[%]	št. nespr.	[%]	št. spr.	[%]
01m	01m_av	219	56,6	168	43,4	164	50,1	163	49,8	383	53,6	331	46,4
	01f_lb	89	57,4	66	42,6	44	46,8	50	53,2	133	53,4	116	46,6
	skupaj	308	56,8	234	43,2	208	49,4	213	50,6	516	53,6	447	46,4
02m	01m_av	211	54,5	176	45,5	133	40,7	194	59,3	344	48,2	370	51,8
	01f_lb	87	56,1	68	43,9	46	48,9	48	51,1	133	53,1	116	46,6
	skupaj	298	55,0	244	45,0	179	42,5	242	57,5	477	49,5	486	50,5
03m	01m_av	242	62,5	145	37,5	203	62,1	124	37,9	445	62,3	269	37,7
	01f_lb	84	54,2	71	45,8	41	43,6	53	56,4	125	50,2	124	49,8
	skupaj	326	60,2	216	39,8	244	58,0	177	42,0	570	59,2	393	40,8
01f	01m_av	160	41,3	227	58,7	106	32,4	221	67,6	266	37,3	448	62,7
	01f_lb	60	38,7	95	61,3	28	29,8	66	70,2	88	35,3	161	64,7
	skupaj	220	40,6	322	59,4	134	31,8	287	68,2	354	36,8	609	63,2
02f	01m_av	171	4,19	216	6,81	113	34,6	214	65,4	284	39,8	430	60,2
	01f_lb	72	46,5	83	53,5	43	45,7	51	54,3	115	46,2	134	53,8
	skupaj	243	44,8	299	55,2	156	37,0	265	63,0	399	41,4	564	58,6
Povprečje	01m_av	200,6	51,8	186,4	48,2	143,8	44,0	183,2	56,0	344,4	48,2	369,6	51,8
	01f_lb	78,4	50,6	76,6	49,4	40,4	43,0	53,6	57,0	118,8	47,7	130,2	52,3
	skupaj	279	51,5	263	48,5	184,2	43,8	236,8	56,2	463,2	48,1	499,8	51,9

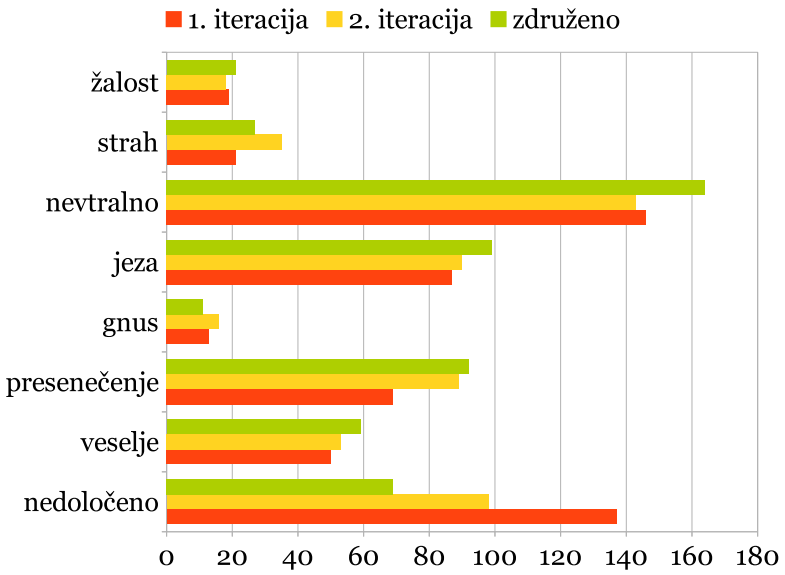
Tabela 4: Pregled ujemanja mnenj označevalcev glede na posamezni podsklop posnetkov. Primerjanje lahko izvedemo samo v podsklopu, kjer je izvedena popolna povezava med posnetki v prvi in drugi iteraciji, medtem ko v sklopu »razdeljene povezanosti« medsebojno primerjamo oznako, ki je bila v prvi iteraciji dodeljena daljšim posnetkom, in oznako, ki je bila v drugi iteraciji dodeljena krajšim.

3.3 Vpliv večinskega mnenja v posamezni iteraciji označevanja

V zbirki EmoLUKS je končna oznaka čustvenega stanja posameznega govorca določena z večinskim mnenjem (ang. majority voting) ocenjevalcev o posameznem posnetku. S takim načinom pripisa končnega čustvenega stanja govorca v posameznem posnetku privzamemo, da je število mnenj označevalcev, ki se strinjajo glede neke oznake, bliže realnemu čustvenemu stanju govorca v posnetku kot v primeru nestrinjanja. Očitno je, da tovrstni pristop lahko privede tudi do nesoglasja označevalcev. V primeru nesoglasja smo definirali novo kategorijo »nedoločeno čustveno stanje govorca«. Ta zajema posnetke posameznika, ki jih na podlagi večinskega mnenja označevalcev ni bilo mogoče določiti. Kot smo opisali v razdelku 2.2, smo pri označevanju poleg čustvenih kategorij označevalcem ponudili v presojo tudi stanje »nič-od-tega«. Ta oznaka opredeljuje bolj kompleksna čustvena stanja govorca, ki niso zajeta v sistemizaciji, ki jo obravnavamo pri označevanju čustvenih stanj govorca. Zaradi manjšega števila tovrstnih oznak, ki so bile pridobljene pri označevanju, te posnetke prav tako preslikamo v novo ustvarjeno kategorijo »nedoločeno čustveno stanje govorca«.

Zasnova označevanja čustev iz posnetkov, pridobljenih iz radijskih iger, nam tudi tokrat narekuje združevanje mnenj posameznih podsklopov posnetkov na način, ki smo ga opisali. To neposredno vpliva na izračun večinskega mnenja označevalcev pri posameznem posnetku. Pri podsklopu posnetkov, ki pomeni polno povezavo med prvo in drugo iteracijo označevanja, smo mnenja ocenjevalcev, pridobljena tako v prvi kot v drugi iteraciji, združili in nato pri vseh desetih mnenjih določili večinsko mnenje, ki pripiše končno čustveno oznako. Slika 1 prikazuje količino posnetkov z enakim večinskim mnenjem v posamezni čustveni kategoriji. Večinsko mnenje posnetkov, ki so bili označeni v prvi (rdeča barva) in drugi iteraciji (rumena barva) označevanja, je določeno na podlagi mnenj petih označevalcev. Združena večinska mnenja pa so bila bila pridobljena z združevanjem mnenj označevalcev v prvi in drugi iteraciji označevanja. Tako je združeno večinsko mnenje (zelena barva) določeno na podlagi desetih mnenj označevalcev. Na Sliki 1 opazimo, da so pridobljena mnenja ocenjeval-

cev v drugi iteraciji označevanja znatno zmanjšala količino posnetkov, ki pripadajo »nedoločeni čustveni kategoriji«. Pri združenih mnenjih ocenjevalcev pa se količina še naprej zmanjšuje za več kot polovico vseh posnetkov iz kategorije »nedoločeno«, označenih v prvi iteraciji.



Slika 1: Analiza vpliva večinskega mnenja za podsklop posnetkov s polno povezanostjo med prvo in drugo iteracijo označevanja glede na zastopanost posnetkov v posamezni čustveni kategoriji.

Pri podsklopu posnetkov »razdeljene povezanosti« lahko na podoben način predstavimo končne oznake posnetkov v prvi in drugi iteraciji označevanja. Končno čustveno oznako posameznega posnetka lahko pripišemo na podlagi pridobljenih mnenj v prvi in drugi iteraciji označevanja skupaj. Slika 2 prikazuje količino označenih posnetkov, ki pripadajo določeni čustveni kategoriji na podlagi večinskega mnenja ocenjevalcev za podsklop posnetkov z imenom »razdeljena povezanost«.

Količina posnetkov, ki je bila označena v prvi (rdeča barva) in drugi (rumena barva) iteraciji označevanja, pa je bila določena s pomočjo večinskega mnenja petih označevalcev. Na Sliki 2 opazimo, da je količina posnetkov znatno manjša

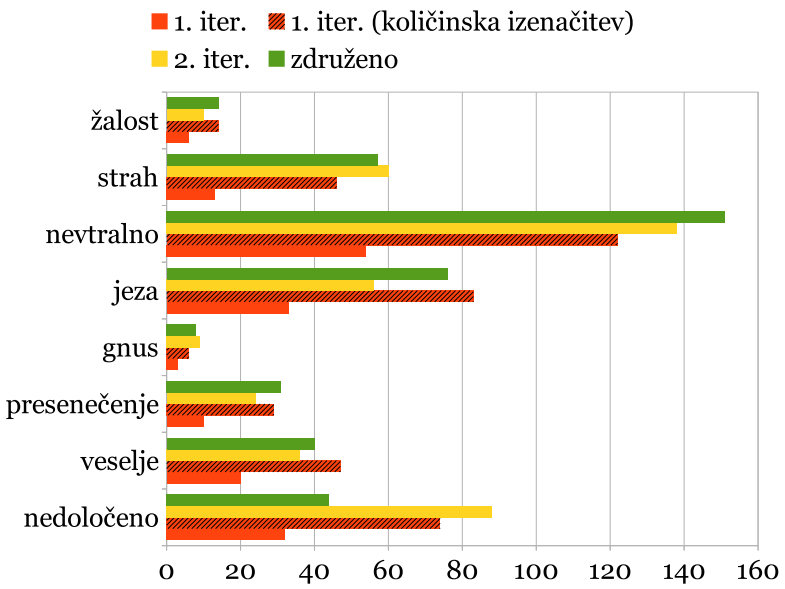
od količine posnetkov v drugi iteraciji (421), saj smo posnetkom v drugi iteraciji določili drugačno ločilno mejo in jih zato lahko predstavimo tudi kot razčlembe (172) posnetkov, ki so jih ocenjevalci označevali v prvi iteraciji označevanja. Da bi ponudili realen vpogled tudi pri združevanju mnenj ocenjevalcev, sliko razširimo s količino razrezanih posnetkov in s pripisanimi oznakami večinskega mnenja v prvi iteraciji. Tovrstno zastopanost posnetkov v posamezni čustveni kategoriji predstavimo s stolpci rdeče barve s poševnimi črtami in jo na Sliki 2 imenujemo 1. iter. (količinska izenačitev).

Z združevanjem mnenj oznak v prvi iteraciji (količinska izenačitev) in drugi iteraciji smo izračunali večinsko mnenje označevalcev, oznake pa pripisali posnetkom, zajetim v zbirko EmoLUKS. Količina posnetkov, ki smo jim določili večinsko mnenje s pomočjo združevanja mnenj ocenjevalcev v prvi in drugi iteraciji, je ponazorjena z zeleno barvo. Posnetki v skupini združeno zavzemajo identične posnetke, kot so jih ocenjevalci označevali v drugi iteraciji, čustvene oznake pa so odraz večinskega mnenja ocenjevalcev, določenega v prvi in drugi iteraciji skupaj.

Iz Slike 2 razberemo, da se pri izračunu večinskega mnenja pri posnetkih druge iteracije poveča število posnetkov, ki pripadajo čustvenim kategorijam strah, nevtralno in gnus. Pri čustvenih kategorijah žalost, jeza, presenečenje in veselje pa se zmanjša. Prav tako se zastopanost posnetkov v drugi iteraciji označevanja, ki jim ne moremo določiti čustvenega stanja na podlagi večinskega mnenja, poveča. Novo porazdelitev zastopanosti posnetkov v posamezni kategoriji lahko pripišemo novi ločilni meji posnetkov. Čeprav ne moremo preveriti, ali so večinske oznake posameznega posnetka s tem načinom označevanja bliže realni čustveni oznaki govorca, vseeno potrdimo, da nova ločilna meja vnaša v označeno gradivo nove zastopanosti posnetkov čustvenih stanj. To nakazuje, da so v radijskih igrah v posameznih povedih tudi večkrat izražena čustvena stanja govorca, ki se v daljših povedih medsebojno prepletajo.

Če združimo oznake ocenjevalcev pri posnetkih, ki so bili označeni v prvi in drugi iteraciji, ter nato izračunamo večinsko mnenje (stolpec, izražen z zeleno

barvo), pridobimo nove zastopanosti posnetkov v posamezni čustveni kategoriji. Iz Slike 2 razberemo, da se količina posnetkov z nedoločeno čustveno oznako pri združenem večinskem mnenju znatno zmanjša v primerjavi s prvo in tudi z drugo iteracijo označevanja. Če primerjamo prvo iteracijo označevanja, se količina posnetkov, ki pripadajo večini čustveni kategorij, poveča. Izjemi sta le jeza in veselje. Prav tako se večina čustvenih kategorij v primerjavi z drugo iteracijo poveča, tokrat sta izjemi le gnus in strah.



Slika 2: Analiza vpliva večinskega mnenja za podsklop posnetkov z razdeljeno povezanostjo med prvo in drugo iteracijo označevanja glede na zastopanost posnetkov v posamezni čustveni kategoriji

Označeno zbirko EmoLUKS pridobimo z izračunom večinskega mnenja v skupinah posnetkov »polna povezanost« in »razdeljena povezanost« pri vseh mnenjih ocenjevalcev v prvi in drugi iteraciji. V končno označeno zbirko seveda vključimo tudi posnetke, ki so bili označeni v prvi ali drugi iteraciji in med seboj nimajo nikakršne odvisnosti, s katero bi lahko združili mnenja ocenjevalcev. Vsem posnetkom podsklopa »brez povezave«, (397 v prvi in 25 v drugi iteraciji), določimo končne oznake na podlagi le petih ocenjevalcev.

3.4 Pregled označenega govornega gradiva v zbirki EmoLUKS

Za zaključek tega poglavja predstavimo v Tabeli 5 v celoti označeno govorno gradivo izbrane govorke in govorca. Deleži označenega gradiva so prikazani kot odstotek zastopanosti posnetkov v posamezni čustveni kategoriji. Zaradi primerjave in skladno s prejšnjimi razdelki predstavimo tudi deleže zastopanosti v posamezni iteraciji označevanja. Označevalci so označili čustvena stanja govorcev, njihovo večinsko mnenje pa lahko razdelimo v osem kategorij. Kot smo že nakazali na Slikah 1 in 2, se tudi pri analizi celotne zbirke EmoLUKS opazi znatno zmanjšanje oznak posnetkov, ki predstavljajo nedoločeno čustveno stanje govorca, na 11 odstotkov. V primerjavi z analizo celotne prve iteracije, kjer smo razpolagali z 18 odstotki tovrstnih oznak posnetkov, in hkrati celotne druge iteracije, kjer ni bilo mogoče določiti čustvene oznake govorca v 20 odstotkih primerov posnetkov, lahko rečemo, da je način združevanja mnenj označevalcev v prvi in drugi iteraciji v podatkovno zbriko EmoLUKS znatno pripomogel k boljši razporeditvi označenih posnetkov v realne čustvene kategorije. Z uporabo metode smiselnega združevanja mnenj označevalcev in/ali posnetkov, označenih v prvi in/ali drugi iteraciji označevanja, smo povečali relativno zastopanost vseh čustvenih kategorij. Čeprav se v Tabeli 5 opazi relativno zmanjšanje količine posnetkov v čustveni kategoriji gnus, naj omenimo, da je absolutna zastopanost kategorije gnus večja kot v obeh iteracijah označevanja. Enake interpretacije pa, žal, ne moremo zagotoviti, če opazujemo skupno trajanje posnetkov. Zbirka EmoLUKS po združevanju in razčlenjevanju, predvsem pa po dodatnih popravkih razčlenitve vsebuje le za 52 sekund več gradiva, kot ga vsebujejo posnetki v prvi iteraciji označevanja, pri govorki 01f_lb celo 12 sekund manj kot pri prvi iteraciji.

	Govorec	Število posnetkov	Trajanje	Delež oznak večinskega mnenja čust. stanj govorca [%]							
				ves.	pres.	gnus	jeza	nev.	str.	žal.	ned.
1. iteracija	01m_av	762	1:01:29	8,5	11,0	1,2	14,5	36,5	5,4	4,9	18,1
	01f_lj	348	14:56	11,8	14,9	4,0	28,5	11,2	8,1	2,3	19,3
	skupaj	1110	1:16:24	9,6	12,3	2,1	18,8	28,6	6,2	4,1	18,5
2. iteracija	01m_av	733	49:12	8,6	9,7	1,4	12,6	35,1	9,4	3,1	20,2
	01f_lj	255	8:27	10,6	17,3	6,3	22,0	13,0	11,0	2,4	17,7
	skupaj	988	57:39	9,2	11,6	2,6	15,0	29,4	9,8	2,9	19,5
EmoLUKS	01m_av	975	1:02:31	8,72	11,6	1,0	15,7	39,7	7,8	4,1	11,4
	01f_lj	410	14:44	12,4	16,8	4,2	27,6	13,2	11,0	3,9	11,0
	skupaj	1385	1:17:16	9,8	13,1	2,0	19,2	31,8	8,7	4,0	11,3

Tabela 5: Pregled deležev označenih posnetkov s čustvenim govorom glede na večinsko mnenje označevalcev za prvo in drugo iteracijo ter združena mnenja prve in druge iteracije v podatkovni zbirki EmoLUKS. Tabela obsega vse mogoče označene kategorije čustvenih stanj, ki so bile označevalcem ponujene v procesu označevanja. Zaradi preglednosti uporabljamo krajšave, ki jih na tem mestu opišemo tudi s polnimi imeni. ves. - veselje, pres. - presenečenje, nev. - nevtralno, str. - strah, žal. - žalost in ned. - nedoločeno.

4 VREDNOTENJE PODATKOVNE ZBIRKE S SAMODEJNIM SISTEMOM ZA RAZPOZNAVANJE ČUSTVENIH STANJ GOVORCA

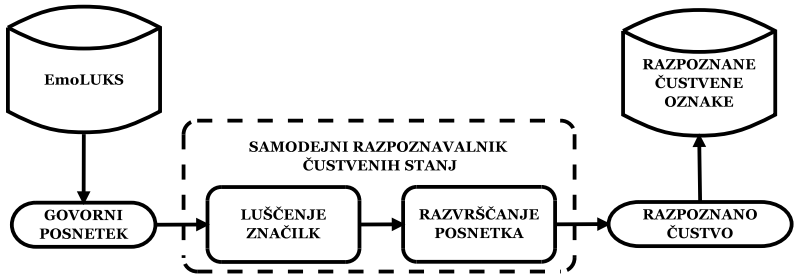
S predstavljenim načinom vrednotenja skušamo odgovoriti na vprašanje, ali posnetki čustvenega govora s pripadajočimi paralingvističnimi oznakami, zbrani v zbirki EmoLUKS, vsebujejo dovolj informacij za samodejno razpoznavo čustvenih stanj govorca. Vrednotenje podatkovne zbirke EmoLUKS izvedemo s pomočjo uveljavljenega sistema za razpoznavanje čustvenih stanj.[1]

4.1 Metodologija vrednotenja zbirke EmoLUKS s samodejnimi razpoznavalniki čustvenih stanj

V tem razdelku najprej opišemo uporabljene algoritme in načine pridobivanja rezultatov, s katerimi lahko ovrednotimo samodejni sistem za potrebe razpoznavanja čustvenih stanj govorca.

4.1.1 SPLOŠNI SISTEM ZA RAZPOZNAVANJE ČUSTVENIH STANJ GOVORCA

Za razvoj sistema za samodejno razpoznavanje čustvenih stanj govorca moramo najprej zagotoviti določeno količino posnetkov čustvenega govora, s katerimi najprej naučimo in nato na preostalem sklopu posnetkov vrednotimo udejanjeni sistem. Splošno zgradbo sistema za razpoznavanje čustvenih stanj iz govora prikazuje Slika 3.



Slika 3: Shematični prikaz uporabe sistema za razpoznavanje čustvenih stanj na podlagi govora pri analizi podatkovne zbirke EmoLUKS.

[1]Uporabljen sistem je bil predstavljen kot referenčni sistem na tekmovanju s področja razpoznavanja čustvenih stanj govorca na konferenci Interspeech 2009 z naslovom »Interspeech 2009 Emotion Challenge« (Schuller idr. 2009a).

Prvi korak, luščenje značilk (Pavešić 2012), je osredotočen na zmanjšanje količine informacij v izvornem posnetku čustvenega govora, pri čemer ohranimo čim večji delež uporabnih informacij, ki so specifične za posamezno kategorijo čustvenih stanj. Rezultat prvega koraka so vektorji značilk. Drugi korak je samodejno razvrščanje čustvenega stanja govorca, ko značilkam posnetka samodejno pripišemo oznako čustva.

4.1.2 IZBOR ZNAČILK ZA VREDNOTENJE PODATKOVNE ZBIRKE EMOLUKS

Če želimo vrednotiti zbirko EmoLUKS in rezultat vrednotenja primerjati z drugimi zbirkami čustvenega govora, moramo v sistem razpoznavanja čustvenih stanj vključiti uveljavljene algoritme za izračun značilk. Najbolj uveljavljen pripomoček za luščenje značilk je prosto dostopno programsko orodje OpenSmile (Eyben idr. 2013). Orodje s pomočjo konfiguracijskih datotek omogoča različno izbiro naborov značilk, ki se pogosto uporabljajo pri raziskovanju samodejnega razpoznavanju čustvenih stanj iz govornih ali videosignalov.

Za preverjanje uspešnosti delovanja sistemov za razpoznavanje čustvenih stanj govorca se večkrat za referenco podajajo rezultati sistemov, ki so bili udejanjeni s postopkom luščenja značilk, ki temelji na statističnih funkcionalih. Postopek je predlagal Schuller idr. (2007). Pokazal je, da je mogoče z relativno preprosto statistično analizo vektorje kratočasovnih akustičnih značilk nekega posnetka predstaviti v strnjeni obliki in pri tem ohraniti informacijo o čustvenem stanju (Gajšek 2012).

Pri naših preizkusih smo za vrednotenje podatkovne zbirke EmoLUKS uporabili nabor značilk, ki je bil določen kot osnovni nabor pri primerjavi razpoznavalnikov na tekmovanju Interspeech 2009 z imenom »Emotion Challenge« (Schuller idr. 2009a).

4.1.3 RAZVRŠČEVALNIK ČUSTVENIH STANJ GOVORCA

Algoritem za razvrščanje je jedro za samodejno razpoznavanje vzorcev. Široka paleta različnih pristopov in metod, ki omogočajo razvrščanje vhodnih vzorcev za potrebe razpoznavanja čustvenih stanj v govoru, je preizkusilo veliko razi-

skovalcev (Anagnostopoulos idr. 2015). Tudi sami smo preizkusili različne postopke za potrebe razvrščanja čustvenega stanja govorca z večdimenzionalnimi značilkami čustvenega govora ter rezultate predstavili v Justin idr. (2010). Najbolj razširjeni razpoznavalniki čustvenih stanj v govoru temeljijo na umetnih nevronskih mrežah (ang. artificial neural networks, ANN) (Dai idr. 2008), linearni diskriminantni analizi (ang. linear discriminant analysis, LDA) (Batliner idr. 2006), odločitvenih drevesih (ang. decission trees) (Ang idr. 2002) in na podlagi podobnosti verjetja modela mešanic Gaussovih porazdelitev (ang. Gaussiona Mixure Model, GMM) (Schuller idr. 2009a). V praksi pa se je v zadnjem času razširila uporaba metode podpornih vektorjev (ang. support vector machine, SVM) (Schuller idr. 2009b; Chavhan idr. 2010), ki smo jo uporabili tudi mi.

Metoda SVM lahko razvršča vhodne vzorce le v dva razreda. Obravnavana zbirka EmoLUKS pa vsebuje oznake sedmih čustvenih stanj govorca. V ta namen moramo metodo SVM razširiti za uporabo večrazrednih problemov razvrščanja. V Hsu in Lin (2002) najdemo več postopkov, ki omogočajo tovrstno razširitev, vendar zaradi svoje enostavnosti izstopata dve, ki sta tudi najpogosteje uporabljeni. Prva se imenuje »eden proti vsem« (ang. one-versus-all), druga pa »eden proti enemu« (ang. one-versus-one). Pri vrednotenju podatkovne zbirke EmoLUKS smo uporabili postopek »eden proti enemu«. V tem primeru zgradimo binarne razvrščevalnike za vsako kombinacijo dveh posameznih razredov. Testni vzorec razvrstimo z vsakim od razvitih razvrščevalnikov in kot zmagovalni razred določimo tistega, ki je izbran s pomočjo večinskega glasovanja. Testnemu vzorcu pripišemo razred, ki je pri preizkušanju vseh kombinacij razvitih razvrščevalnikov največkrat zmagal.

4.1.4 KRITERIJI USPEŠNOSTI SAMODEJNIH SISTEMOV ZA RAZPOZNAVANJE ČUSTVENEGA STANJA GOVORCA

Uspešnost večrazrednega problema razpoznavanja najlaže merimo z matriko zamenjav med razredi (ang. confusion matrix), ki je predstavljana v Tabeli 6. V matriki so C_j oznake posameznih čustvenih razredov, n_{ij} število vzorcev iz ra-

zreda C_i, ki jih je sistem razpoznal kot razred C_j, N_j število posnetkov v razredu C_j in N število vseh posnetkov. Matrika zamenjav je izhodišče za izpeljavo več različnih mer uspešnosti razvrščanja vzorcev.

Osnovni kriterij uspešnosti razvrščanja je predstavljen kot odstotek vseh pravilno razpoznanih vzorcev proti vsem vzorcem v preizkusu (ang. recognition rate). Izračun zanesljivosti razvrščanja (ZR) je predstavljen v enačbi 8.

		Razred razvrstitve				
		C_1	C_2	$\cdots$	C_K	$\sum$
Razred vzorca	C_1	n_{11}	n_{12}	$\cdots$	n_{1K}	N_1
	C_2	n_{21}	n_{22}	$\cdots$	n_{2K}	N_2
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	C_K	n_{k1}	n_{k2}	$\cdots$	n_{KK}	N_K
	$\sum$					N

Tabela 6: Splošna oblika tabele zamenjav, s pomočjo katere lahko predstavimo rezultate razvrščanja

$$ZR = \frac{1}{N} \sum_{k=1}^{K} n_{kk} * 100\% \tag{8}$$

Ta je zanesljiv kazalec uspešnosti razpoznavalnikov le v primerih enakomerne zastopanosti testnih vzorcev v vseh K razredih razpoznavanja. Če imamo neenakomerne porazdelitve zastopanosti vzorcev v razredih razpoznavanja, pa uporabljamo tudi druge kriterije vrednotenja, ki natančneje opisujejo zmožnosti razvrščanja samodejnega sistema za razpoznavanje vzorcev.

Eden takih je natančnost (ang. precision), ki je definiran kot razmerje med številom pravilno razvrščenih vzorcev v posameznem razredu in številom vseh vzorcev, ki jih je sistem razvrstil v isti razred. Natančnost tako opisuje napako, ki je vzrok nepravilnega razvrščanja vzorcev v določeni razred. Natančnost lahko s pomočjo splošne tabele zamenjav zapišemo v enačbo (9), pri čemer je k indeks posameznega razreda razpoznavanja.

$$\text{natančnost}_k = \frac{n_{kk}}{\sum_{i=1}^{K} n_{ik}} \tag{9}$$

Naslednji kriterij imenujemo priklic (ang. recall). Ta je definiran kot razmerje pravilno razvrščenih vzorcev v določenem razredu proti številu vseh testnih vzorcev, ki pripadajo temu istemu razredu. Priklic opisuje napako, ki je posledica razvrščanja vzorcev točno določenega razreda. Tudi priklic lahko s pomočjo splošne tabele zamenjav v Tabeli 6 zapišemo v enačbo (10), pri čemer je k indeks posameznega razreda razpoznavanja.

$$\text{priklic}_k = \frac{n_{kk}}{\sum_{i=1}^{K} n_{ki}} = \frac{n_{kk}}{N_k} \tag{10}$$

Opisana kriterija natančno opredeljujeta uspešnost razpoznavalnika posameznega razreda razpoznavanja in sta dobra kazalca, če želimo ugotoviti, katerega izmed razredov razpoznavalnik slabo ali dobro razvršča. Če pa želimo predstaviti uspešnost razpoznavalnika kot celote, se poslužujemo povprečja vseh razredov. Če pri povprečenju priklica ali natančnosti upoštevamo apriorno verjetnost posameznega razreda, torej kvantitativno zastopanost testnih vzorcev v posameznem razredu, potem pridobimo t. i. utežen priklic in/ali natančnost (ang. weighted average precision/recall), ki ga v nadaljevanju označujemo s kratico WA. Rezultat uteženega povprečnega priklica je enak zanesljivosti razvrščanja. Zato raje predpostavimo, da so vsi razredi enako apriori verjetni in uspešnost sistema predstavimo z neuteženim povprečnim priklicem in/ali natančnostjo (ang. unweighted average recall/precision) in ga v nadaljevanju označujemo s kratico UA. Oba opisana kriterija podajata primernejši opis uspešnosti nekega sistema razpoznavanja, saj izločita vpliv močneje zastopanih razredov razpoznavanja. Zbirke čustvenega govora praviloma ne vsebujejo enakega števila posnetkov vsakega čustvenega stanja, temveč večinoma prevladuje nevtralni govor, zato sta se ti dve meri uveljavili kot kriterij primerjave različnih

sistemov, posledično pa tudi primerjave različnih čustvenih podatkovnih zbirk (Schuller idr. 2009a; Gajšek 2012)

4.1.5 NAČIN VREDNOTENJA - NAVZKRIŽNO VREDNOTENJE

Natančnost poročanja o uspešnosti napovedovanja samodejnega sistema pa ni odvisna le od dobre izbire kriterijskih mer, temveč tudi od delitve razpoložljivih podatkov na dve medsebojno izključujoči se množici, namenjeni učenju in preizkušanju. Razdelitev podatkov, ki so na voljo, poteka po naključnem ključu. Za obravnavani problem razpoznavanja čustvenih stanj govorca se večkrat izkaže, da kvantitativna porazdelitev vzorcev, ki pripadajo posameznemu razredu (čustvenih stanj), ni enakomerno porazdeljena. Na splošno si želimo, da je količina vzorcev v učnem in tudi v testnem delu zastopana enakomerno glede na predhodno določeno razmerje, ki omogoča delitev na testni in učni del. Tak način deljenja zasledimo v literaturi kot stratificirano (ang. stratified) delitev zbirke. Tak način omogoča, da imamo tako v učni kot v testni množici v vsakem od razredov razpoznavanja enak predpisani delež vzorcev glede na celotno zbirko.

Zaupanje v pridobljene rezultate lahko povečamo, če imamo na voljo veliko vzorcev v učni in tudi v testni množici. Pri zbirkah, ki vsebujejo premajhno število vzorcev, da bi lahko uspešno naučili sistem in ga pozneje tudi zanesljivo ovrednotili, uporabimo postopek navzkrižnega vrednotenja (ang. cross validation). Pri tem razdelimo vse vzorce, ki jih imamo na voljo za obdelavo, v K enako velikih delov. Nato za vsak posamezni del k učimo sistem na vseh vzorcih, ki niso v tem delu. Sistem vrednotimo s k-tim delom. Celoten postopek ponovimo K-krat. Po končanem vrednotenju vsakega razvitega sistema dobimo skupno matriko zamenjav, ki zajema razvrstitev vseh vzorcev v zbirki. Obenem vedno razvijemo sistem medsebojno izključujoče se testne in učne množice vzorcev. S takim načinom zagotovimo uravnoteženo vrednotenje in hkrati pridobimo največjo možno testno množico. Če smo pri navzkrižnem vrednotenju pozorni tudi na ohranitev enakih porazdelitev vzorcev, takemu načinu pravimo stratificirano navzkrižno vrednotenje (ang. stratifired cross validation).

4.2 Preizkusi in rezultati vrednotenja zbirke EmoLUKS

V nadaljevanju predstavimo pridobljene rezultate s samodejnimi razpoznavalniki čustvenih stanj govorca pri posnetkih, ki so jim bila pripisana čustvena stanja (govorcu 01m_av in govorki 01f_lb). Pri vrednotenju smo zavrgli vse označeno gradivo, ki ga s pomočjo večinskega menja označevalcev nismo mogli razporediti v eno izmed čustvenih kategorij. Rezultate razvrščanja primerjamo med prvo in drugo iteracijo označevanja ter poročamo o rezultatu razpoznavanja za celotno zbirko EmoLUKS. Ta predstavlja združene rezultate označevanja, kot je bilo predstavljeno v razdelku 2.2. Uspešnost razpozavanja predstavimo v obliki povprečnega uteženega in neuteženega priklica in natančnosti. Kriteriji so podrobneje opisani v razdelku 4.1. Opozoriti je treba, da zaradi različne količine vhodnih posnetkov, ki so bili označeni v prvi in drugi iteraciji, pridobljeni rezultati medsebojno niso absolutno primerljivi. Vseeno pa lahko nakazujejo trend izboljšanja ali poslabšanja verodostojnosti pripisanih oznak tako v prvi kot v drugi iteraciji označevanja podatkovne zbirke EmoLUKS.

Rezultate predstavljamo najprej za dvorazredni problem razpoznavanja vzbujenega in nevtralnega čustvenega stanja govorca, pozneje pa tudi za sedemrazredni problem razpoznavanja kategorij čustvenih stanj govorca. Vsako različico sistema za razpoznavanje čustvenih stanj razvijemo z vsemi mogočimi kombinacijimi podsklopov podatkov glede na identiteto vsakega govorca v zbirki EmoLUKS posebej in tudi za primer združenih posnetkov govorke in govorca. Pri razvoju razpoznavalnikov s podatki ene izmed opazovanih iteraciji vedno pazimo, da so razpoznavalniki razviti na enak način in z enakimi razvrstitvami podatkov v učno in testno množico. Vsakokrat razpoznavalnike vrednotimo s stratificiranim navzkrižnim vrednotenjem.

Za lažje vrednotenje razpoznavalnikov čustvenih stanj smo pripravili program, ki s pomočjo orodja WEKA (Hall idr. 2009) omogoča hitro in preprosto vrednotenje in realizacijo razpoznavalnikov čustvenega stanja govorca.

4.2.1 DVORAZREDNI PROBLEM RAZPOZNAVANJA ČUSTVENIH STANJ GOVORCA

Samodejno razpoznavanje čustvenih stanj iz govora najprej opredelimo kot dvorazredni problem razvrščanja. Vse posnetke, ki so v zbirki EmoLUKS označeni z enim izmed šestih osnovnih čustvenih stanj, preslikamo v skupen razred, ki ga imenujemo »vzbujeno« čustveno stanje. Kot nakazuje že ime, nova oznaka razreda razvrščanja vključuje vse posnetke govora, kjer nam je s pomočjo večinskega mnenja uspelo določiti čustveno stanje. V drugi razred razvrščanja pa uvrstimo vse posnetke, ki so označeni z nevtralnim čustvenim stanjem govorca. S takim načinom pripravimo podatkovno zbirko, ki omogoča graditev dvorazrednega razpoznavanja čustvenih stanj govorca. Razpoznavalnik v tem primeru razvrsti vhodni posnetek v enega izmed dveh razredov razpoznavanja, kot nevtralni govor ali pa kot govor, ki vsebuje eno izmed čustev govorca.

Kadar uporabimo enako identiteto govorca v učnih in testnih podatkih, govorimo o vrednotenju razpoznavalnikov, ki so od govorca odvisni. V našem primeru skušamo predvsem ovrednotiti, ali so pripisane oznake v zbirki EmoLUKS primerne tudi za samodejno razpoznavanje, pri čemer imamo na voljo le označen govor enega govorca in ene govorke. Zato je tak postopek vrednotenja razpoznavalnikov edino smiseln.

Tabela 7 prikazuje vrednotenje udejanjenih razpoznavalnikov. V njej je z odebeljenim besedilom označena zmagovalna vrednost posameznega podsklopa vhodnih podatkov. Tabela 7 tudi omogoča primerjavo rezultatov glede na različne iteracije označevanja zbirke čustvenega govora EmoLUKS. Iz rezultatov je razvidno, da z razpoznavalnikom, ki je bil naučen na posnetkih, označenih v drugi iteraciji, močno izboljšamo rezultat razpoznavanja v primerjavi z vrednotenjem razpoznavalnika, razvitega na podlagi oznak iz prve iteracije. Natančen pregled vrednosti neuteženega povprečnega priklica (UAR) pokaže, da so posnetki, zajeti v drugi iteraciji označevanja, povečali uspešnost razpoznavanja za moškega govorca. Poročanje o absolutnem napredku, žal, zaradi različne količine vhodnih podatkov ni mogoče, očiten pa je enak trend tudi pri rezultatih za žensko govorko.

Govorec	Iteracija	Nabor značilk IS09 dvorazred. SVM [%]			
		UAR	**WAR**	**UAP**	**WAP**
01m_av	1. iteracija	65,2	65,7	65,3	65,7
	2. iteracija	71,8	72,1	71,7	72,2
	EmoLUKS	70,1	70,2	70,0	70,4
01f_lb	1. iteracija	54,5	79,0	54,8	78,3
	2. iteracija	58,0	79,0	58,0	79,0
	EmoLUKS	54,3	79,4	55,4	77,2
Skupaj	1. iteracija	67,9	71,2	68,2	70,9
	2. iteracija	71,2	73,6	71,5	73,4
	EmoLUKS	71,8	74,4	72,1	74,2

Tabela 7: Primerjava uspešnosti razpoznavanja dveh čustvenih stanj (nevtralno in vzbujeno) govorca s stratificiranim 5-kratnim navzkrižnim vrednotenjem razvitih razpoznavalnikov in uporabo nabora značilk IS09. Uspešnost podajamo s kriteriji neuteženega povprečnega (UA) in uteženega povprečnega (WA) priklica (R) in natančnosti (P).

4.2.2 RAZPOZNAVANJE VSEH SEDMIH OZNAČENIH ČUSTVENIH STANJ

Za vrednotenje vseh označenih čustvenih stanj v zbirki EmoLUKS smo prav tako najprej razdelili podatke na učne in testne podmnožice za stratificirano navzkrižno vrednotenje. Količino uporabljenega gradiva za sedemrazredno razvrščanje čustvenih stanj govorca zaradi preglednosti predstavljamo ločeno glede na uporabljeno gradivo pri posamezni iteraciji in nazadnje v združeni zbirki čustvenega govora EmoLUKS. Protokol vrednotenja tudi v tem primeru predvideva preverjanje razvitih od govorca odvisnih razpoznavalnikov.

V Tabeli 8 pa so prikazani rezultati vrednotenja sistemov razvrščanja za vsa označena stanja govorca z uporabo nabora značilk IS09.

Na tem mestu lahko opozorimo, da količina vhodnih podatkov pri razvoju tovrstnih sistemov vrednotenja vpliva na končne rezultate vrednotenja. Opozorimo lahko na najmanj zastopan razred razvrščanja gnus. Za ta čustveni razred pridobimo v vseh razvitih sistemih razvrščanja slabe rezultate, ti pa so seveda zajeti tudi v obravnavnih tabelah, ki vrednotijo razpoznavalnike v celoti za vse mogoče označene razrede v zbirki. O težavni nalogi razvrščanja razreda gnus pričajo re-

Govorec	Iteracija	Nabor značilk IS09 sedemraz. SVM [%]			
		UAR	WAR	UAP	WAP
01m_av	1. iteracija	26,6	47,3	29,5	44,7
	2. iteracija	30,2	49,1	33,8	46,7
	EmoLUKS	27,1	49,6	30,1	46,1
01f_lb	1. iter.	26,7	37,0	27,1	34,9
	2. iteracija	35,5	44,8	34,5	42,8
	EmoLUKS	32,0	44,4	30,0	41,2
Skupaj	1. iteracija	29,0	46,4	30,6	42,7
	2. iteracija	32,0	48,3	33,3	45,4
	EmoLUKS	30,2	47,8	34,1	44,9

Tabela 8: Primerjava uspešnosti razpoznavanja sedmih čustvenih stanj govorca s stratificiranim 5-kratnim navzkrižnim vrednotenjem razvitih razpoznavalnikov in z uporabo nabora značilk IS09. Uspešnost podajamo s kriteriji neuteženega povprečnega (UA) ter uteženega povprečnega (WA) priklica (R) in natančnosti (P).

zultati pri vsakem razvitem sistemu razvrščanja čustvenih stanj. Če opazujemo priklic obravnavanega čustvenega razreda pri vsakem od razvitih sistemov posebej, ugotovimo, da smo pri moškem govorcu vedno pridobili vrednost 0 (pravilno ni bil razvrščen noben vzorec), za primer ženske govorke pa največ 0,06 (pravilno je bil razvrščen le en vzorec). Tako je čustveni razred gnus v zbirki EmoLUKS najslabše zastopano in hkrati tudi najmanj kakovostno označeno govorno gradivo, kar smo že pokazali v razdelku 3.3. Prav tako lahko rečemo, da obravnavani čustveni razred ni primeren za razvoj samodejnih sistemov za razvrščanje od govorca odvisnih čustvenih stanj. Ker smo pri vrednotenju razredov vključili tudi ta problematični razred, vpliv nezmožnosti razvrščanja razreda gnus slabša povprečne rezultate razvrščanja sedemrazrednih razpoznavalnikov čustvenih stanj govorca.

5 SKLEP

V prispevku smo opisali dosedanja prizadevanja pri graditvi slovenske zbirke čustvenega govora iz radijskih iger EmoLUKS. Posebno pozornost smo namenili vrednotenju kakovosti označenih posnetkov govora in predstavitvi izboljšav, ki

smo jih pridobili s pomočjo dvostopenjskega označevanja čustvenih stanj govorca. S predlaganim načinom smo zmanjšali delež govornega gradiva, ki mu na podlagi večinskega mnenja ni bilo mogoče določiti končne čustvene oznake. Ker so bili k ponovnemu označevanju povabljeni isti označevalci, lahko v tem prispevku opazujemo tudi konsistentnost označevalcev pri ponovno označenih posnetkih med prvo in drugo iteracijo označevanja. Izkaže se, da so označevalci v povprečju kar v 51,9 odstotka primerih spremenili svojo mnenje. Takšen podatek lahko povezujemo z neprepričljivim izražanjem čustvenih stanj igralcev iz nabora izbranih čustvenih oznak in seveda tudi z izjemno subjektivnim dojemanjem čustvenih stanj posameznika.

Izračun ujemanja označenih posnetkov s pomočjo Randolphovega načina izračuna κ_{free} in časovno uteženega κ_{tw} koeficienta z vrednostmi med 0,3 in 0,4 lahko opredelimo po Landis in Kochovi karakterizacijski lestvici kot pošteno ujemanje. To dejstvo nakazuje, da kljub težavni nalogi označevanja čustvenega stanja v govoru označeno gradivo vsebuje tudi dobro izražena čustvena stanja v posnetkih govorca in govorke.

V prispevku označeno gradivo vrednotimo s pomočjo uveljavljenega samodejnega postopka razpoznavanja čustvenih stanj govorca. S takim načinom skušamo potrditi konsistentnost postavljenih oznak čustvenega stanja govorca, ki smo jih pridobili na podlagi subjektivnih mnenj označevalcev s pomočjo večinskega mnenja.

Čeprav rezultati samodejnega razpoznavanja čustev in analize ujemanja oznak označevalcev med posameznima iteracijama ter tudi končno določeno zbirko EmoLUKS niso absolutno primerljivi, lahko opazimo trend izboljšave, ki ga pridobimo z oznakami v drugi iteraciji označevanja. S tem lahko tudi potrdimo, da smo s ponovnim označevanjem dela govornega gradiva pridobili bolje označene posnetke, ki odražajo boljše približke k dejanskim čustvenim stanjem govorcev.

Po našem prepričanju so radijske igre primerna izbira za graditev čustvenih govornih zbirk, saj v njih skušajo igralci samo glasovno poslušalcu predstaviti prostor in tudi čas, v katerem se prepletajo dialogi in monologi raznolikih tematik.

Prepričljivost in intenziteta predstavitve sta močno povezana tudi z jasno izraženimi čustvenimi stanji likov, ki jih igralci skušajo z vživeto, hkrati pa tudi premišljeno interpretacijo približati poslušalcu. Zato so čustvena stanja govorcev v takem gradivu nedvomno močno zastopana. Po našem prepričanju tovrstni posnetki odražajo boljši približek k realnim čustvenim stanjem kot pri čustvenih govornih zbirkah, zajetih z namensko interpretacijo vnaprej pripravljenih povedi. Treba pa je tudi poudariti, da je graditev tovrstne zbirke časovno potratnejša.

Zbirka EmoLUKS trenutno obsega označen čustven govor ene govorke in enega govorca. Podatkovno zbirko bomo skušali v prihodnosti nadgraditi z dodatnim označevanjem preostalega gradiva drugih govorcev in govork, pri katerih želimo uvesti tudi označevanje obsežnejšega nabora čustvenih oznak s pripadajočo intenziteto.

LITERATURA

Anagnostopoulos, C.-N., Iliou, T. in Giannoukos, I. (2015): Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, *43* (2): 155–177.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E. in Stolcke, A. (2002): Prosody-based automatic detection of annoyance and frustration in human-computer dialog. V J. H. L. Hansen in B. L. Pellom (ur.): *7th international conference on spoken language processing*: 2037–2040. Denver: ISCA.

Ayadi, M. E., Kamel, M. S. in Karray, F. (2011): Survey on speech emotion recognition: features, classification schemes, and databases. *Pattern Recognition*, *44* (3): 572–587.

Barras, C., Geoffrois, E., Wu, Z. in Liberman, M. (2001): Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, *33* (1–2): 5–22.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N. in Kessous, L. (2006): Combining efforts for improving automatic classification of emotional user states. V T. Erjavec in J. Gros (ur.): *Jezikovne tehnologije, IS–LTC 2006*: 240–245. Ljubljana: Inštitut Jožef Stefan.

Chavhan, Y., Dhore, M. in Pallavi, Y. (2010): Speech emotion recognition using support vector machine. *International Journal of Computer Applications*, *1* (20): 6–9.

Cohen, J. (1960): A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cornelius, R. R. (1996): *The science of emotion: research and tradition in the psychology of emotions*. Prentice-Hall, Inc.

Cornelius, R. R. (2000): Theoretical approaches to emotion. *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*: 3–11. ISCA. Newcastle.

Cowie, R. in Cornelius, R. R. (2003): Describing the emotional states that are expressed in speech. *Speech Communication*, *40* (1–2): 5–32.

Dai, K., Fell, H. J. in MacAuslan, J. (2008): Recognizing emotion in speech using neural networks. V R. Merrell (ur.): *V proceedings of the IASTED International Conference on Telehealth/Assistive Technologies*: 31–36. Anaheim: ACTA Press Anaheim.

Douglas-Cowie, E., Campbell, N., Cowie, R. in Roach, P. (2003): Emotional speech: towards a new generation of databases. *Speech Communication*, *40* (1–2): 33–60.

Eyben, F., Weninger, F., Groß, F. in Schuller, B. (2013): Recent developments in OpenSMILE, the Munich open-source multimedia feature extractor. *V international conference on Multimedia*: 835–838. Barcelona: ACM.

Fleiss, J. L. (1971): Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76* (5): 378.

Gajšek, R., Štruc, V., Mihelič, F., Podlesek, A., Komidar, L., Sočan, G. in Bajec, B. (2009a): Multi-modal emotional database: AvID. *Informatica (Ljubljana)*, *33* (1): 101–106.

Gajšek, R. (2012): *Samodejno razpoznavanje čustvenega stanja na podlagi govora*. Doktorska disertacija, Univerza v Ljubljani, Fakulteta za elektrotehniko.

Gajšek, R., Štruc, V., Vesnicer, B., Podlesek, A., Komidar, L. in Mihelič, F. (2009b): Analysis and Assessment of AvID: Multi-Modal Emotional Database. V V. Matoušek in P. Mautner (ur.): *Text, speech and dialogue*: Zv. 5729: 266–273. Springer Berlin Heidelberg.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. in Witten, I. H. (2009): The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11* (1): 10–18.

Howe, J. (2006): The rise of crowdsourcing. *Wired magazine*, *14* (6): 1–4.

Hozjan, V., Kačič, Z., Moreno, A., Bonafonte, A. in Nogueiras, A. (2002): Interface Databases: Design and Collection of a Multilingual Emotional Speech Database. *V proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*: 2019–2023. Las Palmas de Gran Canaria: ELRA.

Hsu, C.-W. in Lin, C.-J. (2002): A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, *13* (2): 415–425.

Juang, B. H. in Rabiner, L. R. (1991): Hidden markov models for speech recognition. *Technometrics*, *33* (3): 251–272.

Justin, T. (2016): *Umetno tvorjenje čustvenega slovenskega govora z uporabo prikritih Markovovih modelov*. Doktorska disertacija, Univerza v Ljubljani, Fakulteta za elektrotehniko.

Justin, T., Gajsek, R., Štruc, V. in Dobrišek, S. (2010): Comparison of different classification methods for emotion recognition. V N. Bogunović in S. Ribarić (ur.): *MIPRO, 2010 proceedings of the 33rd international convention*: 700–703. Opatija: IEEE.

Justin, T., Mihelič, F. in Žibert, J. (2014): Razvoj zbirke slovenskega emocionalnega govora iz radijskih iger - EmoLUKS. V T. Erjavec in J. Ž. Gros (ur.): *Zbornk 9. konferenceca Jezikovne tehnologije*: 157–162. Ljubljana: Inštitut Jožef Stefan.

Koolagudi, S. G. in Rao, K. S. (2012): Emotion recognition from speech: a review. *International Journal of Speech Technology*, *15* (2): 99–117.

Landis, J. R. in Koch, G. G. (1977): The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Mihelič, F., Gros, J., Dobrišek, S., Žibert, J. in Pavešić, N. (2003): Spoken language resources at LUKS of the University of Ljubljana. *International Journal of Speech Technology*, *6* (3): 221–232.

Pavešić, N. (2012): *Razpoznavanje vzorcev: uvod v analizo in razumevanje vidnih in slušnih signalov*. Ljubljana: Založba FE in FRI.

Plutchnik, R. (1962): *The emotions: facts, theories, and a new model*. Random House.

Randolph, J. J. (2005): Free-marginal multirater kappa (multirater k [free]): an alternative to fleiss' fixed-marginal multirater kappa. *V prispeveku predstavljenem na: Joensuu University Learning and Instruction Symposium 2005*. Joensuu: ERIC.

Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N. in Kessous, L. (2007): The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. *V 8th annual conference of the international speech communication association*: 2253–2256. Antwerp: ISCA.

Schuller, B., Steidl, S. in Batliner, A. (2009a): The INTERSPEECH 2009 emotion challenge. V B. Schuller, S. Steidl in A. Batliner (ur.): *V 10th annual conference of the international speech communication association*: 312–315. Brighton: ISCA.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C. in Narayanan, S. (2013): Paralinguistics in speech and language – State-of-the-art and the challenge. *Computer Speech & Language*, *27* (1): 4–39.

Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G. in Wendemuth, A. (2009b): Acoustic emotion recognition: A benchmark comparison of performances. *V Automatic Speech Recognition & Understanding, 2009*: 552–557. IEEE.

Vesnicer, B. in Mihelič, F. (2004): Evaluation of the Slovenian HMM-based speech synthesis system. V P. Sojka, I. Kopeček in K. Pala (ur.): *Text, speech and dialogue*: Zv. 3206: 513–520. Springer Berlin Heidelberg.

Development and Evaluation of the Emotional Slovenian Speech Database – EmoLUKS

The paper describes development of the Slovenian emotional speech database for its primary use in speech synthesis. We also explore the potential of additional annotation for extending it for the use in emotion recognition tasks. The paper focus in methodology for annotating para lingual speaker information on the example of annotating speaker emotions in Slovenian radio dramas. Emotional speech database EmoLUKS was built from speech material of 17 Slovenian radio dramas. We obtained them from the national radio-and-television station (RTV Slovenia), which were given to the universities disposal with an academic license for processing and annotating the audio material. The utterances of one male and one female speaker were transcribed, segmented and then annotated with emotional states. The annotation of the emotional states was conducted in two stages with our own web-based application for crowd sourcing. Annotating assessments in different time periods with same five volunteers allows us to compare the obtained annotator's decisions, therefore we report about annotator's decisions consistency. Based on annotators majority vote of each annotated utterance we label speech material and join it to emotional speech database named EmoLUKS. The material currently consists of 1385 recordings from one male (975 recordings) and one female (410 recordings) speaker and contains labeled emotional speech with a total duration of around 1 hour and 15 minutes. The paper presents the two-stage annotation process used to label the data and demonstrates the usefulness of used annotation methodology. We evaluate the consistency of the annotated speech material with the speaker dependent automatic emotion recognition system. The reported results are presented with the un-weighted as well as weighted average recalls and precisions for 2-class and 7-class recognition experiments. Results additionally confirms our presumption, that emotional speech database despite its complexity includes also clearly expressed emotional speaker states.

Keywords: emotional speech database, emotion recognition