Razvoj zbirke slovenskega emocionalnega govora iz radijskih iger - EmoLUKS

Tadej Justin[1], **France Mihelič**[1], **Janez Žibert**[2]

[1] Univerza v Ljubljani,Fakulteta za elektrotehniko, LUKS, Tržaška 25, 1000 Ljubljana
{tadej.justin, france.mihelic}@fe.uni-lj.si
[2] Univerza na Primorskem, Fakulteta za matematiko, naravoslovje in informacijske tehnologije, Glagoljaška 8, 6000 Koper
janez.zibert@upr.si

Povzetek

V tem delu predstavljamo gradnjo slovenske zbirke emocionalnega govora za namen umetnega tvorjenja govora in razpoznavanja emocionalnih stanj govorca. V prispevku se osredotočamo na opis razvite metodologije in razvoj programske opreme za označevanje paralingvistične informacije v govoru na primeru označevanja emocionalnih stanj v slovenskih radijskih igrah. Govorno zbirko in programsko opremo za množično označevanje smo v celoti zasnovali v Laboratoriju za umetno zaznavanje, sisteme in kibernetiko na Fakulteti za elektrotehniko v Ljubljani. Zbirka vsebuje govorne zvočne signale, ki so del sedemnajstih radijskih iger, s katerimi razpolagamo z licenco za akademsko uporabo. Za namen testiranja razvite aplikacije namenjene množičnemu označevanju posnetkov v tem prispevku poročamo o označeni zbirki ene govorke in enega govorca. Označevanje emocionalnih stanj je označilo pet prostovoljcev. S pomočjo razvite spletne aplikacije, ki temelji na sistemu za urejanje vsebin CMS Plone je bilo označenih 1110 posnetkov. Dodatno v prispevku poskušamo predstaviti problematiko povezano z označevanjem govornih zvočnih posnetkov na primeru množičnega označevanja emocionalnih stanj v govoru iz govornih posnetkov radijskih iger in poročamo o ujemanju oznak označevalcev.

Development of emotional Slovenian speech database based on radio drama - EmoLUKS

In this article we present the development of the Slovenian emotional speech databases developed for purposes of speech synthesis and automatic emotion recognition. The main focus in this article is about the development of methodology and software used to label the paralinguistic information from speech. The design of the database and development of the software for crowd-sourcing was produced and developed at the Laboratory of Artificial Perception, Systems and Cybernetics at Faculty of Electrical Engineering of Ljubljana. The curently annotated database consists of speech signals extracted from 17 radio dramas, with the academic licence for processing and annotating the audio signals authorized by from RTV Slovenia. For purposes of testing the developed crowd-sourcing software we focused in labelling emotional speakers states of one male and one female speaker. The emotional labels were annotated using the developed web based application with five volunteers. In this article we present the implementation of web based application for crowd-sourcing based on CMS Plone and annotating procedure which results in emotional speech database consisting of 1110 recordings. We additionally focus in the problems of annotating the speech corpora in the crowd-sourcing environment for annotating the paralinguistic informations from speech and on the example of the annotated database we report about the obtained annotations based on annotators majority vote.

1. Uvod

Dolgoletni razvoj sistemov za razpoznavanje in umetno tvorjenje govora (sintezo govora) je dodobra izpopolnil metode in principe obeh področji. Vseeno pa še vedno ostaja prostor za razvoj sistemov, ki omogočajo razpoznavanje ter tvorjenje paralingvističnih stanj govorca (Yamashita, 2013). V zadnjem desetletju je veliko pozornosti namenjene raziskavam in razvoju sistemov, ki omogočajo razpoznavanje emocionalnih stanj govorca (Ayadi et al., 2011), kakor tudi sistemov namenjenih tvorbi emocionalnega govora (Krstulovic, 2007). Dan danes so na globalnem trgu že prisotne naprave ter aplikacije, ki omogočajo interakcijo človek-stroj (ang. human-computer-interaction, HCI) preko govora. Tovrstne aplikacije so močno odvisne od jezika, kar je tudi eden od razlogov, da je opravljanje tovrstnih aplikaciji v Slovenskem jeziku mnogokrat zapostavljeno. V želji, da bi aplikacijam omogočili bolj naravno tvorbo umetnega govora ter pripomogli k raziskovanju avtomatskega razpoznavanja emocionalnih stanj, je nedvomno eden izmed prvih korakov k raziskovanju tovrstnih aplikativnih sistemov gradnja od jezika odvisne govorne podatkovne zbirke z dodatnimi paralingvistični oznakami govorca.

Do danes je bilo razvitih veliko tuje jezičnih govornih zbirk, ki skušajo zajeti tudi paralingvistična stanja govorca (Schuller et al., 2013). Tovrstna stanja se v literaturi opisujejo kot stanje govorca, katero se ne da opisati z lingvističnimi ali fonetičnimi oznakami. Paralingvistična stanja so lahko izražena v govoru kot na primer zastrupljenost, razpoloženje, zanimanje, emocionalno stanje, itd.. Načrtovanje izgradnje tovrstnih podatkovnih zbirk zahteva zahtevno interdisciplinarno sodelovanje. Eden izmed pomembnejših dejavnikov predstavlja prav opredelitev paralingvističnih oznak, kjer je nujno potreben ekspert začrtanega področja uporabe podatkovne zbirke. Tako na primer označevanje emocionalnih stanj v govoru, predstavlja težavno nalogo, saj trenutno ne razpolagamo z splošno uveljavljeno metodologijo opisovanja emocionalnih stanj. V takem primeru se velikokrat raziskovalci zatečejo k utečenim postopkom izgradnje govornih podatkovnih zbirk po zgledih v svetovni literaturi, ki opredeljuje natančne opise emocionalnih stanj v govoru in se glede na potrebe raziskovanja močno razlikujejo. V splošnem lahko potrebne opise emocionalnih stanj delimo glede na zastavljen cilj uporabe. V literaturi (Cowie and Cornelius, 2003) lahko zasledimo tipične raziskovalne potrebe, ki narekujejo smernice k izgradnji govornih emocionalnih podatkovnih zbirk in so predvsem osredotočene k raziskovalnemu cilju. Gradnjo takih podatkov-

nih zbirk največkrat usmerijo raziskovalni cilji ki strmijo k raziskovanju teoretskega ozadja emocionalnih stanj v govoru in so v večini primerov psihološke ali biološke narave. Na drugi strani so zastopani tudi cilji, ki strmijo k uporabi tovrstnih podatkovnih zbirk v aplikativne namene.

V slovenskem prostoru so do danes prisotne dve govorni zbirki emocionalnega govora, katerih opise najdemo v (Gajsek et al., 2009) in (Hozjan et al., 2002). Prva predstavlja multi-modalno zbirko spontanih emocionalnih stanj in je njena uporaba za namen sinteze govora zelo omejena. Druga predstavlja del večjezične govorne zbirke Interface, ki je dostopna pod komercialno licenco.

V tem prispevku se osredotočamo na gradnjo emocionalne govorne podatkovne zbirke za aplikativno uporabo v namen sinteze slovenskega emocionalnega govora. Predstaviti želimo dosedanje delo in probleme s katerimi se srečujemo oblikovalci tovrstnih podatkovnih zbirk. Govorna zbirka, ki jo predstavljamo v tem prispevku je izdelana preko že zajetih govornih posnetkov slovenskih radijskih iger.

Prispevek delimo na štiri poglavja, kjer v metodologiji predstavimo potrebne aplikacije za izgradnjo emocionalnih podatkovnih zbirk. Nadaljujemo z rezultati, ki predstavijo označen emocionalni govorni material ter hkrati posvetimo posebno pozornost ujemanju mnenj označevalcev. V naslednjem poglavju skušamo povzeti probleme pri gradnji tovrstne zbirke. V zaključku prestavimo nadalje delo ter komentiramo označeni del podatkovne zbirke za ciljno uporabo v sistemu za slovensko emocionalno umetno tvorjenje govora.

2. Metodologija

Dan danes je uspešnost avtomatskih sistemov, ki uporabljajo algoritme s področja umetne inteligence ter strojnega učenja, močno odvisna od velikega števila vzorcev (učna množica), ki so na razpolago za učenje modela. Uspešnost se določi s pomočjo postopkov za evalvacijo ter s pomočjo vzorcev, ki so na razpolago za testiranje (testna množica). V evalvacijskem postopku največkrat primerjamo rezultate udejanjenega sistema ter označbe vzorcev, ki so pripisane testnim vzorcem. S pridobljeno uspešnostjo lahko rečemo, da udejanjeni sistem lahko dobro ali slabo opravlja svojo nalogo z uspešnostjo tudi na naravnih vzorcih, ki niso del podatkovne zbirke na podlagi katere smo ga razvili. Dobra strategija pri izdelavi podatkovne zbirke namenjene tako učenju ter testiranju sistemov za specifično nalogo je torej ključnega pomena za udejanjanje splošnih sistemov za določeno nalogo.

V primeru pridobivanja govornih podatkovnih zbirk, namenjenih razpoznavanju in/ali tvorjenju umetnega govora, lahko opazimo, da so močno odvisne od jezika. Želja vseh razvijalcev s tega področja je pridobiti tako podatkovno zbirko, ki zajema čim več jezikovnih prvin, tako iz pisane besede kot tudi iz glasoslovja. Glavna razlika med govornimi podatkovnimi zbirkami namenjenimi umetnemu tvorjenju govora ali razpoznavanju govora, je v številu govorcev zajetih v podatkovni zbirki. V prvem primeru si v splošnem želimo razpolagati z obsežno zbirko enega govorca, ki vsebuje čim večje število različnih posnetkov. V drugem primeru pa si želimo razpolagati z zbirko čim večjega števila različnih govorcev. Tovrstne podatkovne zbirke ponavadi vsebujejo posnetke enakih v naprej predvidenih stavkov prebranih s strani več govorcev. S takimi strategijami v prvem primeru pridobimo dovolj raznolik in čim boljši približek govorjene besede (jezika) posameznega govorca. V drugem primeru, pa si želimo razviti čim bolj robusten model, ki omogoča dobro razpoznavanje čim večjega števila uporabnikov.

Razvijalci emocionalnih govornih podatkovnih zbirk, ki so namenjene za aplikativno rabo v avtomatskih sistemih za umetno tvorjenje govora ali razpoznavanje emocionalnih stanj govorca, se velikokrat poslužujejo dveh strategij, ki omogočata zajem zbirke. Prvi predstavlja snemanje govorne zbirke s pomočjo profesionalnih bralcev, ki so zmožni posnemati emocionalna stanja. Take zbirke so posnete z pomočjo vnaprej pripravljenih povedi, ki so izbrane iz obsežnejših zbirk besedil in v njihovi celoti skušajo zadostiti fonetični porazdelitvi osnovnih enot posameznega jezika. V drugem primeru pa razvoj zbirke zajema pridobivanje že posnetkih govornih segmentov, ki jih je potrebno točno in natančno prepisati ter v primeru večjega števila govorcev na posnetku dodatno časovno določiti identiteto govorca v posameznem posnetku. V obeh primerih je potrebno govorne posnetke označiti z vnaprej predvidenimi emocionalnimi oznakami. V zadnjem času se za to nalogo velikokrat najame označevalce, katerih večinsko mnenje določa končno oznako posameznega posnetka. V primeru podatkovne zbirke EmoLUKS predstavljajo označeni signali igrana emocionalna stanja govorcev, saj so radijske igre posnete z poklicnimi igralci.

2.1. Transkripcija in segmentacija govornih posnetkov

S pomočjo RTV Slovenija smo pridobili radijske igre, ki so bile v večini posnete v profesionalnem studiu radia Slovenija. Vsako radijsko igro smo transkribirali ter razčlenili glede na identiteto govorca. V poteku segmentacije in transkribiranja smo želeli označiti predvsem čisti govor, zato smo vzporedno označevali tudi nejezikovne prvine, ki so večkrat del radijskih iger. To se v večini glasba v ozadju, različni šumi ter raznovrstni dodatni zvočni efekti. Poleg tega nismo pozabili tudi ostale nejezikovne prvine govorca, kot so vdih, cmokanje, stokanje, jok in smeh.

Za potrebe prepisov in razčlenitve glede na govorca smo uporabili programa Transcriber (Barras et al., 2001). Orodje omogoča hitro in učinkovito razčlenjevanje govornih signalov glede na govorca, njihovo transkripcijo ter označevanje ne jezikovnih elementov v posnetku. Posnetke smo razčlenili tudi glede na zaključene stavčne enote. S takim pristopom smo pridobili nabor posnetkov, ki niso predolgi in hkrati nudijo dovolj konteksta za označevanje paralingvistične informacije v govoru.

Z orodjem Transciber smo označili 17 posnetkov radijskih iger v približnem skupnem časovnem obsegu 12 ur in 50 minut. Tabela 1 na strani 3 prikazuje količino transkribiranega in označenega materiala.

2.2. Definicija emocionalnih stanj

Vsakršno raziskovalno delo, ki posega na področje čustev, potrebuje najprej definicijo, kaj čustvo je oziroma, kaj

št.	Naslov radijske igre	Trajanje
1	Penzion Evropa	0:48:03,56
2	Angleško poletje	0:57:55,69
3	V Sieni nekega deževna 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	Nas 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 trajanj celotnih radijskih iger

s širokega področja analize čustev pri človeku bo središče preučevanja. Razlikovanja v teoretskih predpostavkah na katerih temelji teorija o čustvih (Cornelius, 1996), pričajo o razlikovanju tolmačenja čustev. V literaturi se pojavljajo štirje različni pogledi na čustveno stanje (Cornelius, 2000). Imenujemo jih Darvinistični pogled, pogled po Jamesu, kognitivistični pogled in socialno-konstruktivistični pogled. Preko različnih pogledov na čustvena stanja se posledično uporabljajo tudi različni modeli, ki opisujejo relacije med različnimi čustvenimi kategorijami. Osnovna predpostavka, na kateri temeljijo razmejitve med čustvenimi kategorijami pri vseh modelih, je da so razlike med opaženimi čustvenimi doživljaji znotraj ene kategorije manjše od razlik med tistimi iz različnih kategorij. Pregled modelov čustev je predstavljen v (Cornelius, 1996), kjer avtor predstavlja modele znotraj štirih glavnih skupin in jih imenuje prostorski, diskretni, pomenski in komponenti modeli.

V tem prispevku se avtorji osredotočamo na diskretizacijo čustvenih stanj po Darwinovem pogledu na čustvena stanja. Tako se osredotočamo na predpostavko, da obstaja nekaj osnovnih čustev, iz katere so se razvili osnovni diskretni modeli čustvenih kategorij. Tak pristop je tudi eden izmed najpopularnejših predstavitev prostora čustvenih stanj. Na tak način smo diskretizirali osnovna čustvena stanja v naslednje kategorije, ki smo jih uporabili za označevanje posnetkov radijskih iger: žalost, veselje, gnus, jeza, strah, presenečenje in nevtralno.

Označevanje emocionalnih stanj v govoru poteka s pomočjo ekspertnega znanja. Govornim posnetkom lahko pripiše oznako ekspert za dano področje. V zadnjem času se večkrat uporablja nabor označevalcev, ki podajo svoje mnenje o posameznem posnetku. S takim naborom mnenj lahko bolj posplošeno določimo oznako posnetku. Ker je označevanje govornih posnetkov velikokrat dolgotrajen proces, se za označevanje podatkovnih zbirk večkrat uporablja spletne aplikacije, ki omogočajo hkratno označevanje večjega števila označevalcev ter hkrati ponujajo označevalcem svobodno izbiro časovnega okvira označevanja. V literaturi tovrstni pristop zasledimo pod pojmom množično izvajanje (ang. croud-sourcing) (Howe, 2006).

2.2.1. Razvoj in opis aplikacije za označevanje zvočnih in video posnetkov

Po pregledu literature ter ogledom dostopnih spletnih aplikacij, ki nudijo označevanje govornih posnetkov, smo se odločili, da nobena v taki meri ne izpolnjuje pogojem, ki bi morali biti zadoščeni, da lahko enostavno ponudimo našim prostovoljnim označevalcem kvalitetno in hitro označevanje. Zato smo se odločili izdelati spletno aplikacijo namenjeno označevanju zvočnih ali video posnetkov. K taki odločitvi nas je napeljalo tudi dejstvo, da lahko razpolagamo in obdelujemo tovrstne podatke samo za akademske potrebe. Zato bojazen, da ob prenosu na spletno mesto večjih razsežnosti ter posledični kraji intelektualne lastnine, ki nam je bila zaupana v varstvo s strani Radia Slovenija zadostimo le v primeru, če omogočimo gostovanje posnetkov tovrstne aplikacije na lastnih spletnih strežnih. Z uporabo sodobnih tehnologij ter nenehnim varovanjem in nadzorovanjem strežniškega sistema se lahko kraji takih podatkov izognemo, vendar le v primeru, če razpolagamo z popolnim nadzorom nad aplikacijo in strežnikom.

Spletno aplikacijo smo razvili s pomočjo odprto kodnih prosto dostopnih programov. Spletna aplikacijo je bila razvita kot dodatek k sistemu za urejanje vsebin (ang. Content Management System, CMS) Plone verzije 4.3.3 [1]. Odprto kodni sistem CMS Plone je razvit na podlagi programskega ogrodja za urejanje vsebin Zope [2]. Izbira takega osnovnega sistema za razvoj aplikacije sloni na dejstvih, da je sistem Plone eden izmed spletnih CMS, ki se ponaša z eno boljših sledi o zapisih varnostnih popravkov [3] ter je zaradi tega uvrščen v skupino najbolj varnih CMS. Poleg tega ima že vgrajen način za delo z delovnimi tokovi (ang. workflows), ki so nujno potrebni za enostavno urejanje nad pravicami za ogledovanje, urejanje in ustvarjanje spletnih vsebin. Hkrati ponuja razvoj dodatkov, ki jih lahko razvijalec implementira in namesti v že obstoječi sistem.

Spletno aplikacija sestoji iz uredniških in uporabniških strani. Tako uredniki kot tudi uporabniki (ocenjevalci) se morajo v sistem prijaviti z uporabniškim imenom in geslom. Ker je označevanje zvočnih ali video posnetkov lahko dolgotrajen proces, smo s takim pristopom zagotovili označevalcem možnost označevanja v daljšem časovnem obdobju. Hkrati smo onemogočili naključnim obiskovalcem dostop do občutljivih posnetkov. Spletna aplikacija namenjena označevanju zvočnih ali video posnetkov je dostopna na http://emo.luks.fe.uni-lj.si.

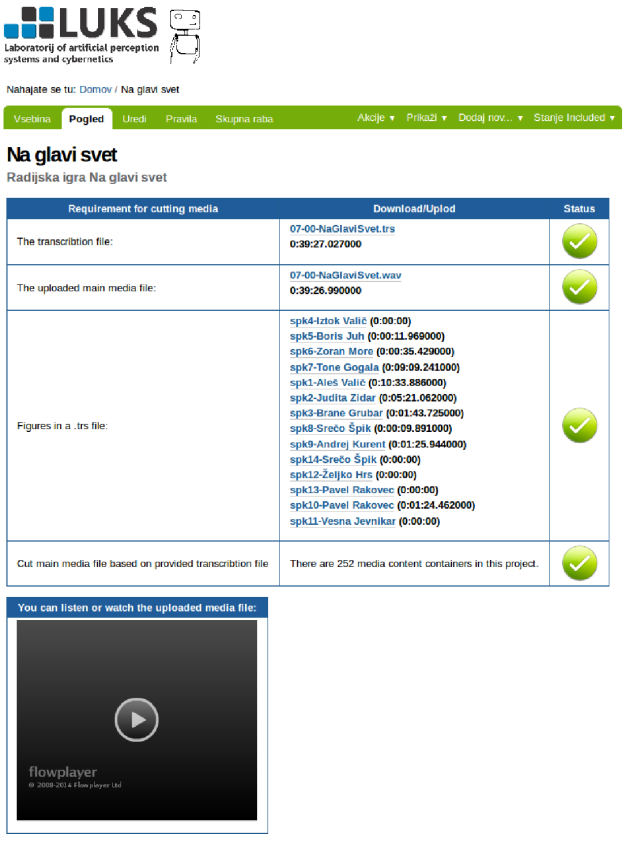
2.2.2. Uredniške strani

Poleg preprostega označevanje paralingvistične informacije v zvočnih ali video posnetkih, ki jo nudi spletna aplikacija smo pripravili tudi uredniški vmesnik, ki omogoča enostavno in hitro izdelavo projekta označeva-

[1] http://plone.org
[2] http://zope.org
[3] http://cve.mitre.org/

nja. Uredniku je omogočen enostaven prenos obsežnejše zvočne ali video datoteko s pripadajočo datoteko v tekstovni obliki, ki vsebuje potrebne zapise o segmentaciji in transkribciji. Trenutno podprt format datoteke za segmentacijo je format Transcriber XML. Spletna aplikacija avtomatsko razreže posnetek na manjše zaključene posnetke, ki so časovno označeni v datoteki za segmentiranje. Določi jim tudi identiteto govorca ter jasno razdeli posnetke s čistim govorom ter posnetki, ki vsebujejo poleg govora tudi druge nejezikovne prvine. Aplikacija poskrbi tudi za pravilen format prikaza zvočnih ali video posnetkov v različnih spletnih brskalnikih. Slika 1, prikazuje uspešen uvoz potrebnih podatkov na spletni strežnik.

Uredniku spletne aplikacije je po uspešnem uvozu podatkov omogočena enostavna izdelava projekta označevanja. Urednik z vnosom zahtevanih parametrov ustvari nalogo namenjeno označevanju. V tem delu ima možnost vključiti katere večje enote posnetkov bodo na voljo za označevanje. V našem primeru so to radijske igre. Uredniku je na tem mestu omogočeno, da lahko na enostaven način vključi v nalogo označevanja le del vseh dostopnih sklopov podatkov na strežniku. Poleg tega lahko urednik enostavno izbira iz nabora vseh identitet govorcev le tiste, za katere meni, da je označevanje smiselno. V tem primeru bodo v nalogo označevanja vključeni le posnetki, ki vključujejo govor vključenih identitet govorcev. K začetni inicializaciji in selekciji govorcev v postopku ustvarjanja naloge označevanja sodi tudi podroben opis naloge označevanja ter opis samega procesa označevanja. Uredniku je omogočeno preko pisane besede, slik in povezav v splet jasno predstaviti označevalcu cilje označevanja in hkrati pregledno opisati in definirati katere paralingvistične oznake so na voljo za označevanje posnetkov. V posebno polje urednik vnese skrajšana imena z nabora opredeljenih paralingvističnih oznak. Ta imena so kasneje uporabnikom v procesu označevanja prikazana kot možna izbira za označbo posnetka. Ker aplikacija zahteva tudi uvoz transkripciji lahko urednik omogoči izpis prepisa zvočnega posnetka. Ker so v večini paralingvistični dejavniki v govoru zaznani preko širšega konteksta aplikacija omogoča poleg prikaza transkribcije samega posnetka tudi izpis širšega nabora transkribciji pred in po posnetkom, ki ga označuje označevalec. Tak pristop označevalcu nudi vpogled v predhodno ter kasnejše dogajanje, ki omogoča označevalcu umestitev posnetka v širši kontekst. V primeru označevanja radijskih iger se to izkaže za koristno, saj so velikokrat prisotni dialogi med dvema govorcema in s tako ponazoritvijo dialoga označevalcu pojasnimo kontekstno dogajanje.



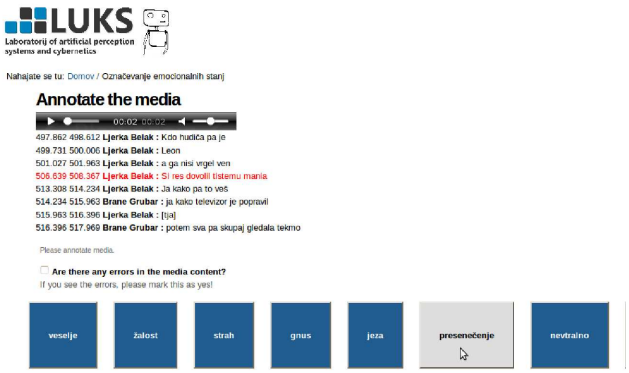
Slika 1: Uredinške strani aplikacije za označevanje, primer uspešnega vnosa podatkov.

Pri dolgotrajnem procesu transkribiranja in segmentiranja zvočnih ali video posnetkov se velikokrat pojavijo napake. Urednik spletne aplikacije ima možnost omogočiti sistem poročanja o napakah. Ta sistem označevalcem omogoča enostavno označbo, da posnetek vsebuje napako. Taki posnetki so uredniku prikazani ločeno, ki jih lahko enostavno popravi s pomočjo urejanja posnetka kar preko spleta.

V primeru razvoja paralingvističnih govornih podatkovnih zbirk iz že vnaprej posnetega govornega materiala so razvijalci primorani material, ki ga hočejo vključiti podatkovno zbirko v celoti pregledati in žal tudi nekaj material, ki ni smiseln ali pa je posledica nenatančne sementacije v postopku transkribiranja in segmentiranja zavreči. Aplikacija zato urednikom ponuja pred objavo procesa označevanja izvedbo testnega protokola označevanja. Uredniku je ponujena možnost identičnega označevanja, ki je kasneje na voljo posameznemu označevalcu. Med potekom testa ima urednik možnost enostavne izključitve posnetka iz procesa označevanja.

2.2.3. Uporabniške strani

Po prijavi v spletni sistem je označevalec najprej seznanjen z nalogo in opisom protokola označevanja zvočnih ali video posnetkov. Z sprejetjem pogojev označevanja lahko označevalec prične z delom. Na zaslonu sem mu avtomatsko predvaja govorni ali video posnetek iz nabora posnetkov, ki so del označevalnega procesa. Vrstni red posnetkov, ki jih označevalec označuje je naključen. Označevalec s klikom na gumb pod posnetkom izbere med možnimi izbirami oznak. Po izbiri oznake se mu na zaslon prikaže naslednji posnetek namenjen označevanju in se avtomatsko predvaja. V kolikor se označevalec zmoti ima vedno možnost popravka zadnjih pet odločitev. Slika 2 prikazuje označevanje posnetka za osnovna emocionalna stanja. V kolikor urednik omogoči dodatne opcije, se pod predvajalnikom posnetka izpiše tudi širši kontekst prepisov, kateremu sledi tudi možnost označbe napake v posnetku ali v transkribciji. Sledijo možne izbire posameznih vnaprej predvidenih emocionalnih stanj, ki so v našem primeru na voljo za označevanje.



Slika 2: Uporabniška stran v postopku označevanja emocionalnih stanj.

Označ.	Povprečen čas odločitve	Št.popravkov odločitev	Skupen čas odločitev
01m	26,04	14 (1,26 %)	08:01:44,08
02m	10,12	6 (0,54 %)	03:07:16,43
03m	15,10	1 (0,09 %)	04:39:27,65
01f	30,51	2 (0,18 %)	09:24:07,88
02f	31,68	10 (0,90 %)	09:46:07,88
Skupaj	**22,64**	**33 (0,59 %)**	**34:58:43,93**

Tabela 2: Označevanje 1110 posnetkov s povprečnim trajanjem 4,13 sekunde iz sklopa 17 radijskih iger petih označevalcev. Prvi stolpec označuje identiteto označevalca ter spol.

Aplikacija ima vgrajen tudi sistem za merjenje časa, ki ga označevalec porabi za odločitev. Ker ima označevalec vedno možnost večkratnega poslušanja enega posnetka lahko z analizo takih podatkov hitro ugotovimo najbolj problematične odločitve. Hkrati nudi kontrolo nad najmanjšim časovnim obsegom, ki ga mora poslušalec nameniti za izvedbo odločitve. Porabljen čas mora biti vedno večji, kot pa čas posnetka predvajanja.

Ko označevalec označi vse predvidene posnetke se proces označevanja zaključi. Takrat je označevalcu ponujen vpogled v porabo časa, ki ga je namenil za označevanje in hkrati vpogled v kratek pregled števila označenih posnetkov.

2.3. Postopek označevanja emocionalnih stanj pri zbirki Emo LUKS

V tem prispevku predstavljamo trenutno označeno delo, ki obsega govorca ženskega in moškega spola. V nabor označevanja smo vključili vse radijske igre. Ocenjevalcem je bilo omogočeno poročanje napak ter tudi izpis širšega konteksta transkribcije posnetka. V označevalnem procesu je bilo vključenih 1110 posnetkov v skupnem časovnem obsegu 1 ure 16 minut in 24 sekund. Povprečni čas trajanja posnetka, ki je vključen v proces označevanja je 4,12 sekunde. V procesu označevanja je sodelovalo pet označevalcev. Trije moški in dve ženski. Vsi označevalci so uporabljali slušalke. Vsak od označevalcev je označil vse posnetke. V tabeli 2 so prikazani povprečne vrednosti časa potrebnega za izvedbo odločitev, število popravkov odločitev označevalca ter skupen efektivni čas označevanja vseh 1110 posnetkov.

Vsak označevalec je lahko izbiral med osnovnimi kategorijami emocionalnih stanj; žalost, veselje, gnus, jeza, strah, presenečenje in nevtralno. Poleg osnovnih stanj govorca je bila dodana tudi kategorija ”nič-od-tega”, ki označevalcu omogoča označbo emocionalnega stanja govorca, ki ni del predpisane sistemizacije kategorij osnovnih emocionalnih stanj v govoru.

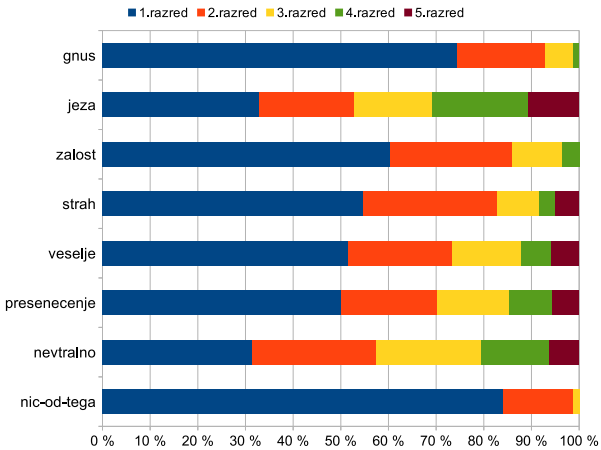
3. Rezultati

Spletna aplikacija omogoča enostaven izvoz vseh označenih podatkov v tekstovno obliko formata ”.csv” (ang. coma separated value). Tak format podpira večina programskih orodji za statistično obdelavo in analizo podatkov.

Ujemanja mnenj označevalcev prikazuje slika 3. Na sliki so prikazani deleži odločitev v petih razredih. Vsak izmed razredov predstavlja razmerje števila mnenj označevalcev o posameznem posnetku v posamezni emocionalni kategoriji proti številu vseh mnenj za posamezno emocionalno kategorijo. Na ta način je v prvem razredu zastopan delež posnetkov za katere je en označevalec podal mnenje o posamezni kategoriji. V drugem razredu so zastopani deleži posnetkov, pri katerih sta se dva označevalca odločila za enako mnenje. Analogija o definiciji posameznih razredov se lahko enostavno razvleče do petega razreda.

Iz slike 3 lahko razberemo, da je ujemanje vseh petih ocenjevalcev o posameznem posnetku (5. razred) zastopano pri označbah jeza, strah, veselje, presenečenje in nevtralno. Delež ujemanja vseh petih ocenjevalcev pri navedenih kategorijah označb se giblje od 4,9% do 10,5%. Ker so deleži popolnega ujemanje premajhni in bi z takim pristopom določevanja oznak posnetkov pridobili le malo število posnetkov, nekatere kategorije pa bi bili primorani celo zavreči se končna oznaka posameznega posnetka določi s pomočjo večinskega mnenja označevalcev (ang. majority voting).

Tabela 3 prikazuje pridobljen material s pomočjo večinske odločitve označevalcev na posameznem posnetku. Tre-



Slika 3: Slika ujemanje mnenj o posameznih posnetkih.

Govorec	Št. posnet.	Trajanje	Deleži oznak čustvenih stanj [%]								
			nev.	jeza	ves.	pres.	žal.	strah	gnus	ned.	nič
01m_av	762	01:01:28,60	36,48	14,44	8,53	11,02	4,86	5,38	1,18	16,67	1,44
01f_lj	348	14:55,55	11,21	28,45	11,78	14,94	2,30	8,05	4,02	17,82	1,44
skupaj	1110	01:16:24,15	28,56	18,83	9,55	12,25	4,05	6,22	2,07	17,03	1,44

Tabela 3: Pregled deležev označenih emocionalnih posnetkov s pomočjo določitve večinskega strinjanja označevalcev.

nutno razpolagamo z oznakami petih označevalcev preko vsega govornega materiala iz 17 radijskih iger, identitete govorca *av* in govorke *lb*.

4. Diskusija

Označevanje emocionalnega stanja govorca se velikokrat izkaže za težavno nalogo, saj ni splošno sprejete definicije kategorij emocionalnih stanj. To dejstvo potrjujejo tudi rezultati, saj se menja označevalcev o posnetkih velikokrat razlikujejo. Z podrobnim pregledom slike 3 na strani 5, lahko ugotovimo, da do popolnega konsenza med označevalci prihaja le v redkih primerih. Vseeno lahko potrdimo da je v petih od sedmih klasifikacij med označenimi oznakami emocionalnih stanj v deležu med približno 5% in 10% procentov prisoten popolni konsenz med mnenji označevalcev. Slednje nakazuje na dejstvo, da so v slovenskih radijskih igrah jasno izražena emocionalna stanja igralcev in da je izbira radijskih iger smiselna za gradnjo zbirke slovenskega emocionalnega govora.

Večinsko mnenje označevalcev prikazuje tabela 3 na strani 6.Iz podatka o deležu nedoločenih oznak nad posnetki opazimo, da je 17% izmed vseh posnetkov, takih katerim ne moremo s pomočjo večinskega odločanja določiti emocionalnega stanja. Izmed vseh nedoločenih posnetkov je 91% takih, katerim sta dva označevalca pripisala eno emocionalno stanje, druga dva pa drugo emocionalno stanje in le 9% takih, katerim je vseh pet označevalcev pripisalo različno emocionalno stanje. Na tem mestu se zdi smiselno za posnetke z nedoločenim stanjem postopek označevanja ponoviti in s tem preveriti konsistentnost označevalcev ter opazovati ali posnetki resnično vsebujejo večdimenzionalna oziroma prepletajoče se emocionalna stanja.

Zasnovana spletna aplikacija se je se je izkazala za uspešno izbiro, ki omogoča hiter in učinkovit način označevanja paralingvističnih stanj v govornih ali video posnetkih tako za razvijalce podatkovne zbirke, kot tudi za označevalce. Vseeno lahko iz tabele 2 na strani 5, opazimo veliko časovno odstopanje potrebno za označevanje med ženskimi in moškimi označevalci. Razloge za to lahko iščemo v dveh dejavnikih. Prvi se nanaša na natančno označevanje z večkratnim poslušanjem govornih posnetkov ter drugi na šibko internetno povezavo, ki lahko vpliva na hitrost prenosa kratkih govornih posnetkov.

5. Zaključek

V prispevku smo predstavili spletno aplikacijo za množično označevanje paralingvistične informacije v govornih ali video posnetkih na primeru označevanja emocionalnih stanj v govoru iz slovenskih radijskih iger za namenom izgradnje govorne zbirke slovenskega emocionalnega govora - EmoLUKS.

Čeprav podatkovna zbirka vsebuje nekoliko manj govornega materiala za posamezno emocionalno stanje, kot smo to pričakovali se vseeno nadejamo, da lahko z uporabo sodobnih pristopov za tvorjenje umetnega govora s pomočjo Prikritih Markovih Model in priročnih postopkov adaptacije akustičnih modelov, ne glede na manjšo količino materiala, ki je natančno označena pripomoremo k večji naravnosti umetnega govora.

6. Literatura

Moataz El Ayadi, Mohamed S. Kamel, in Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572 – 587.

Claude Barras, Edouard Geoffrois, Zhibiao Wu, in Mark Liberman. 2001. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1–2):5 – 22.

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

Randolph R Cornelius. 2000. Theoretical approaches to emotion. V: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

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

Rok Gajsek, Vitomir Struc, France Mihelic, Anja Podlesek, Luka Komidar, Gregor Socan, in Bostjan Bajec. 2009. Multi-modal emotional database: Avid. *Informatica (Slovenia)*, 33(1):101–106.

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

Vladimir Hozjan, Zdravko Kacic, Asuncion Moreno, Antonio Bonafonte, in Albino Nogueiras. 2002. Interface databases: Design and collection of a multilingual emotional speech database. V: *LREC*.

Sacha Krstulovic. 2007. A review of state-of-the-art speech modelling methods for the parameterisation of expressive synthetic speech. Tehnično poročilo, DFKI Deutsches Forschungszentrum für Künstliche Intelligenz.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, in Shrikanth Narayanan. 2013. Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4 – 39.

Yoichi Yamashita. 2013. A review of paralinguistic information processing for natural speech communication. *Acoustical Science and Technology*, 34(2):73–79.