Review article / Pregledni rad
Manuscript received: 2016-04-26
Revised: 2016-05-29
Accepted: 2016-05-31
Pages: 53 - 74

Metodologija estimacije emocionalnih stanja na temelju akustičkih značajki govora

Branimir Dropuljić
IN2data d.o.o.
Zagreb, Hrvatska
branimir.dropuljic@in2data.eu

Sandro Skansi
IN2data d.o.o.
Zagreb, Hrvatska
sandro.skansi@in2data.eu

Leo Mršić
IN2data d.o.o.
Zagreb, Hrvatska
leo.mrsic@in2data.eu

Sažetak: U novije vrijeme se sve veća pažnja posvećuje problematici računalne estimacije emocionalnog stanja iz čovjekovog glasa, prvenstveno u kontekstu razvoja sustava za inteligentnu interakciju između čovjeka i računala. U radu je opisana metodologija estimacije po koracima: izvlačenje akustičkih značajki emocionalnog govora, redukcija prostora značajki te estimacija emocionalnih stanja na temelju neke od metoda strojnog učenja. Emocije se tipično reprezentiraju kao diskretna stanja, poput sreće, ljutnje, straha ili gađenja, ili kao dimenzije, najčešće kao razine ugode i pobuđenosti. Pritom se za raspoznavanje diskretnih emocija koriste klasifikacijske metode, a za estimaciju dimenzijskih veličina emocija regresijske. U radu je dan pregled *state-of-the-art* akustičkih značajki za prepoznavanje emocija te su prikazani rezultati relevantnih radova na ovom području.

Ključne riječi: estimacija emocionalnih stanja, afektivno računarstvo, govorni signal, akustičke značajke, baze emocionalnog govora

UVOD

Mjerljivost emocija u glasu zaintrigirala je znanstvene krugove još 1930-tih godina. M. Cowan je 1936. godine napravio analizu akustičkih glasovnih značajki govornih signala snimljenih tijekom javnih nastupa [5]. Nekoliko godina kasnije, Fairbanks i Pronovost analizirali su širi spektar emocija u govoru [11]. Veliki znanstveni doprinosi nastali su 1980-tih godina, kad su napravljene prve značajnije analize utjecaja emocija na prozodijsku strukturu govora. Posebno su važni radovi Fricka [13] i Scherera [41] iz tog perioda. Razvoj tehnologija kao što su pozitronska emisijska tomografija (engl. *positron emission tomography*, PET), funkcionalna magnetska rezonanca (engl. *functional magnetic resonance imaging*, fMRI) i slične, omogućio je velik napredak u razumijevanju neurološke pozadine procesa nastanka emocija te njihov utjecaj na govor. Pored psihologa i lingvista se u istraživanja na ovom području aktivnije uključuju i neuroznanstvenici te se krug interesa za ovu interdisciplinarnu tematiku počeo širiti.

Značajni doprinosi ostvareni su od strane računalnih znanstvenika, i to razvojem sustava i metoda za raspoznavanje emocija na temelju akustičkih ili lingvističkih značajki govora. Ozbiljnosti ovakvih specifičnih primjena metoda digitalne obrade signala i strojnog učenja svakako pridonosi novo-uspostavljena grana računarstva: afektivno računarstvo (engl. *affective computing*). Ovo interdisciplinarno područje koje uključuje računarsku znanost, psihologiju i kognitivnu znanost [55] svoje začetke temelji na radu Rosalind Picard iz 1995. godine [36].

Brojna istraživanja su pokazala da akustičke značajke vrlo dobro koreliraju s emocionalnim stanjima govornika. Utvrđeno je da se relevantne akustičke značajke mogu dobiti iz fonacijske i artikulacijske faze govora [1], [29], [31], [41], [45] što se uglavnom odnosi na statističku analizu kontura sljedećih govornih mjera tijekom izgovora: fundamentalne frekvencije, tj. osnovne frekvencije titranja glasnica; intenziteta glasa (energije); brzine prolazaka govornog signala kroz nulu (engl. *zero-crossing rate*); parametara vokalnog trakta dobivenih iz spektralne distribucije glasa; te harmonijske strukture glasa. Detaljan pregled relevantnih akustičkih mjera i značajki za prepoznavanje emocija biti će prikazan u nastavku rada.

Estimacija emocionalnih stanja uglavnom se provodi nekom od metoda strojnog učenja. Odabir metode i njezinih parametara u velikoj mjeri ovisi o veličini skupa glasovnih značajki te načinu reprezentacije emocija. U literaturi se emocije najčešće definiraju diskretnim ili dimenzijskim modelima. Za klasifikaciju diskretnih emocionalnih stanja, poput sreće, tuge, straha, itd., često su korišteni modeli s Gaussovim mješavinama (engl. *Gaussian mixture models*, GMM) [33], [45], [60]. Značajni rezultati klasifikacije emocija iz glasa postizani su sa strojevima s potpornim vektorima (engl. *support vector machines*, SVM) [19], [50]. Navedeni modeli emocije analiziraju kao statična stanja. Korištenje skrivenih Markovljevih modela (engl. *hidden Markov models*, HMM) omogućava analizu vremenskog slijeda glasovnih uzoraka te na taj način omogućava praćenje dinamike procesa emocija [35].

Analiza diskretnih emocija je zastupljena u najviše radova iz ovog područja, no u novije vrijeme se sve više autora okreće i prema drugim oblicima reprezentacije emocija, na-

jčešće dimenzijskim modelima ugode i pobuđenosti. To je djelomično povezano s činjenicom da su procesi emocija po svojoj prirodi dinamični i vremenski promjenjivi [26], te ih je kao takve smislenije opisivati numeričkim vrijednostima nego klasama. Wollmer i sur. [59] navode sljedeći argument u korist dimenzijskih modela: »Ljudske emocije su kontinuiran proces i sustav za automatsko prepoznavanje emocija ih mora moći prepoznati kao takve.« Također, noviji radovi se često fokusiraju na razvoj računalnih sustava koji imaju sposobnost rada u stvarnom vremenu te je mogućnost kontinuiranog praćenja emocionalnog stanja govornika od velike važnosti [10].

Glasovne značajke mogu se podijeliti na dvije skupine:

- *lingvističke značajke* – izvučene iz sadržaja govorne poruke u formatu ključnih riječi ili složenijih jezičnih izraza koji predstavljaju određenu informaciju o emocionalnom stanju govornika;
- *akustičke značajke* – mjere varijacije verbalnog iskaza, odnosno načina na koji je nešto izrečeno.

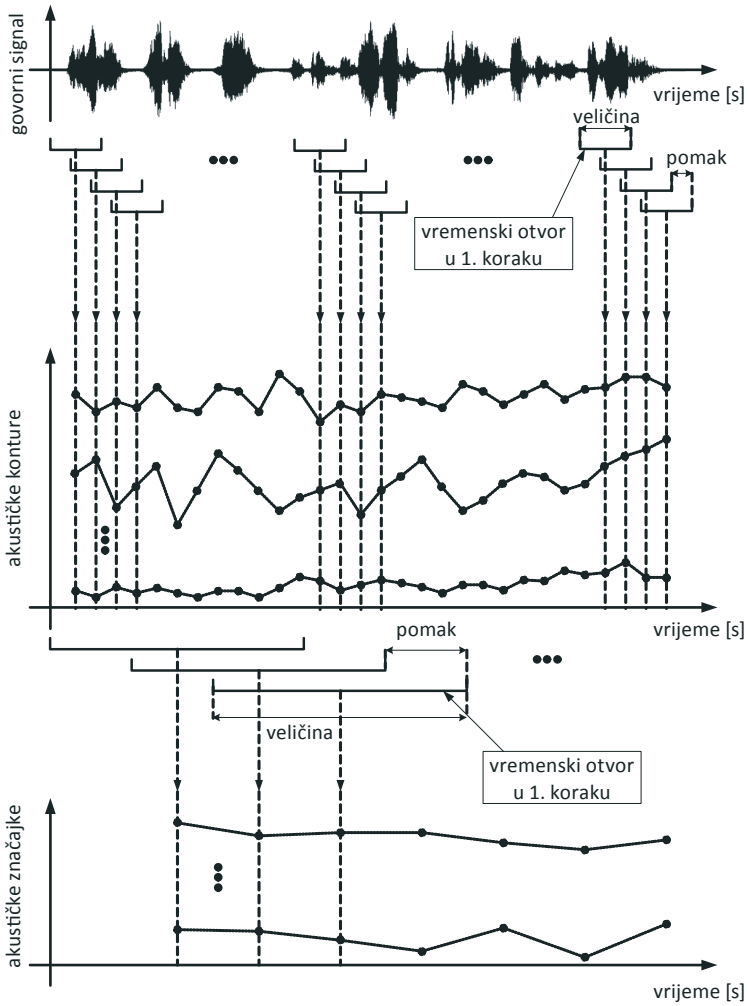
U literaturi su za prepoznavanje emocija češće korištene akustičke značajke. Akustičke značajke su u većoj mjeri univerzalne, to jest ne ovise toliko o jeziku govornika i manje su podložne voljnoj kontroli. U radovima se uglavnom postiže veća točnost korištenjem skupa akustičkih značajki, u usporedbi sa skupom lingvističkih, no neki radovi pokazali su i da se točnost estimacije može povećati koristeći fuziju akustičkih i lingvističkih značajki [10], [24], [45], [46]. Kako te dvije komponente glasa nisu nužno korelirane, sveobuhvatna fuzija glasovnih značajki može pružiti integralnu informaciju o utjecaju emocija na govor.

Ipak, lingvističke značajke su se u nekim primjenama pokazale kao dobro korelirane s emocijama. Tako je detaljnim pregledom područja u [6] ustanovljeno kako su lingvističke značajke bolje korelirane sa složenijim (sekundarnim) emocijama poput ljubomore, ponosa i sličnih, dok su se akustičke značajke očekivano pokazale kao dobro korelirane s primarnim (s evolucijskog stanovišta – arhetipskim) emocijama: srećom, tugom, strahom, ljutnjom, iznenađenjem i gađenjem. S obzirom da su se sekundarne emocije formirale paralelno s razvojem kulture, a time i jezika [7], takvi rezultati imaju uporište. Nadalje, u [10] i [59] je pokazano da u slučaju estimacije dimenzijskih emocija lingvističke značajke bolje koreliraju s ugodom, dok akustičke bolje koreliraju s pobuđenošću.

AKUSTIČKE ZNAČAJKE

Akustičke značajke se u većini slučajeva računaju kroz dva koraka. U prvom se koraku obrađuje govorni signal koji je prikupljen pomoću senzora (mikrofona) te uzorkovan i kvantiziran pomoću analogno/digitalnog (A/D) pretvornika (npr. na zvučnoj kartici računala). Iz takvog se digitalnog signala računaju određene akustičke mjere i to u slijednom procesu, korištenjem vremenskog otvora. Vremenskim se otvorom (Hammingovim, Hannovim, Blackmanovim ili nekim drugim) izdvajaju odsječci govornog signala, koji se dodatno atenuiraju ovisno o odabranom otvoru te se nad takvim odsječkom računaju sve zadane mjere. Ovaj se proces ponavlja duž cijelog govornog signala pomicanjem vremenskog otvora, što rezultira nizom vrijednosti, tj. konturom svake pojedine mjere.

Iz kontura se u idućem koraku, također pomoću vremenskih otvora ali sad s drugačije postavljenim parametrima, računaju akustičke značajke govornog signala. Konačan rezultat ovog postupka za zadani govorni signal može biti vektor značajki ili matrica, odnosno skup kontura značajki, ovisno o dužini trajanja govornog signala i veličini vremenskog otvora. Drugi korak definiran na ovakav način (uz primjenu otvora) nalazi primjenu uglavnom u sustavima koji su prilagođeni za *on-line* način rada (u stvarnom vremenu), dok se kod *off-line* analiza (bez ograničenja stvarnog vremena) uglavnom računa jedan vektor značajki na temelju cjelokupnog izgovora.

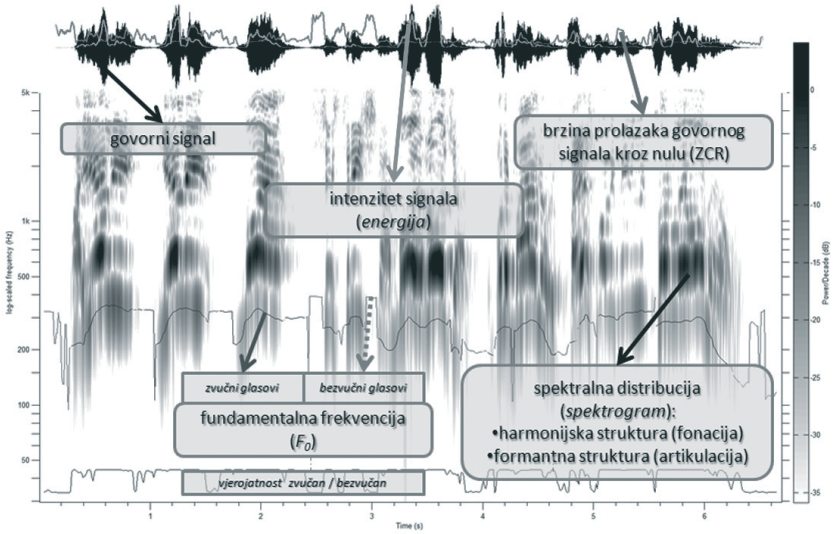


Slika 1: *Postupak izračuna akustičkih značajki iz govornog signala kroz dva koraka uz pomoć vremenskih otvora.*

Postupak izračuna značajki kroz dva koraka ilustriran je na slici 1. Veličina i pomak vremenskog otvora za obradu govornog signala su postavljeni na vrijednosti 25ms, odnos-

no 10ms, što se u praksi pokazalo kao dobar balans između potrebne količine uzoraka za analizu i zadržavanja svojstva kvazi-stacionarnosti govornog signala. Nad ovako definiranim parametrima provodi se takozvana kratkotrajna analiza govora (engl. *short-term speech analysis*). S obzirom da su vremenski otvori uglavnom zvonolikog oblika (s izuzetkom pravokutnog otvora) iznos pomaka se obično definira negdje u intervalu trećine do polovine veličine vremenskog otvora. Na ovaj način su svi uzorci govornog signala reprezentativno zahvaćeni u analizi.

U ovom poglavlju opisane su akustičke značajke emocionalnog govora koje se gore opisanim postupkom računaju iz sljedećih akustičkih mjera glasa: fundamentalne frekvencije, harmonijske strukture, spektralne distribucije, intenziteta, trajanja izgovora i drugih. Ilustracija ovih mjera na primjeru izgovora jedne rečenice napravljena je na slici 2. Iz ovih se akustičkih mjera u većini slučajeva značajke računaju kao statističke mjere poput: ekstrema (minimum, maksimum i sl.), srednje vrijednosti, percentila, momenata, regresijskih koeficijenata i mnogih drugih [48].



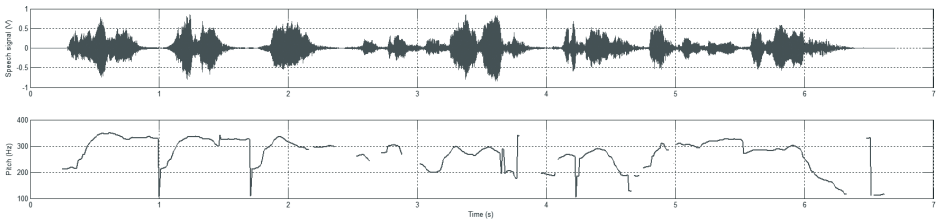
Slika 2: *Prikaz akustičkih mjera iz glasa korištenjem alata Voicebox [61].*

FUNDAMENTALNA FREKVENCIJA

Glasnice titraju pri prolasku zraka iz pluća kod zvučnih glasova, odnosno samoglasnika i zvučnih suglasnika. Kod bezvučnih suglasnika glasnice su u otvorenom položaju i samo propuštaju zrak. Titranjem glasnica formira se harmoničan signal koji putuje dalje kroz vokalni trakt. Takav pobudni signal sastoji se od niza harmonika, gdje osnovni (prvi) harmonik odgovara upravo frekvenciji titranja glasnica, odnosno fundamentalnoj frekvenciji (F_0).

Potrebno je razlikovati termine fundamentalna frekvencija i *pitch*, gdje se *pitch* u striktnom smislu odnosi na slušnu percepciju osnovnog tona u zvuku. Takva percepcija pod-

ložna je utjecaju i ostalih parametara zvuka, poput primjerice glasnoće. Fundamentalna frekvencija je inherentno svojstvo periodičkih signala i može se definirati kao inverz najmanjeg 'pravog' (osnovnog) perioda signala. S obzirom da F_0 nije moguće egzaktno izračunati iz govornog signala, za estimaciju se koriste tzv. *pitch trackeri*, odnosno računalni algoritmi koji *pitch* računaju pomoću brojnih metoda kao što su: autokorelacija ili kros-korelacija, kepstralna analiza, primjena tzv. češljastog filtra (engl. *comb filter*) i drugih [54]. Ilustracija izračuna F_0 konture tijekom izgovora prikazana je na slici 3.



Slika 3: *Primjer govornog signala i estimirane fundamentalne frekvencije. Za estimaciju je korišten RAPT algoritam (engl. Robust Algorithm for Pitch Tracking), u sklopu Voicebox alata [61], predložen u [54].*

Fundamentalna frekvencija se mijenja tijekom prirodnog izgovora, što je posljedica lingvističkih zahtjeva, to jest formiranja naglasaka kod povezivanja fonema u slogove, kao i izmjena zvučnih i bezvučnih intervala unutar riječi i rečenica. Na ovaj način formira se F_0 kontura kroz izgovor, gdje su navedene promjene superponirane na bazičnu F_0 vrijednost govornika. Takva bazična vrijednost varira među govornicima (poznato je da muška populacija u prosjeku ima niži bazični F_0, odnosno »dublji« glas, od ženske populacije), ali i za pojedinog govornika, što ovisi o nizu unutarnjih i vanjskih faktora.

Na bazičnu vrijednost i varijabilnu komponentu F_0 konture u velikoj mjeri utječu i emocije. Brojni radovi su pokazali da je upravo ova komponenta, kao indirektna mjera aktivacija i kontrakcija osjetljivih unutarnjih mišića grkljana, u velikoj mjeri podložna promjenama pod utjecajem emocija i kao takva predstavlja temelj u analizi emocionalnih reakcija iz glasa. U [40] se navodi da je u literaturi ustanovljena vrlo dobra koreliranost F_0 s raznim emocionalnim stanjima, odnosno s pobuđenošću i aktivacijom diskretnih emocija, ali i s psihopatološkim stanjima poput anksioznosti ili raznih stanja napetosti, shizofrenije te depresije.

U literaturi su za prepoznavanje emocija iz F_0 uglavnom korištene statističke značajke poput srednje vrijednosti, minimalne i maksimalne amplitudne vrijednosti na intervalu, raspona između te dvije vršne vrijednosti, varijance (ili standardne devijacije) i percentila (posebice medijana) [1], [2], [41], [43], [45]. Ovakve mjere uglavnom pokazuju dobru povezanost s emocionalnim stanjima, ali ponajprije u smislu statističkog opisa voljnih promjena u izgovoru, to jest fazičnih aktivacija somatskog (voljnog) živčanog sustava (SNS). Srednja vrijednost F_0 konture tijekom izgovora, prikazana kao relativna mjera pre-

ma *baseline* vrijednosti govornika, može poslužiti kao dobar pokazatelj napetosti mišića grkljana, odnosno toničnih aktivacija SNS-a kod izgovora [41].

Slične statističke mjere korištene su i u analizi zasebnih intervala F_0 konture, kao što su intervali rastućih i padajućih nagiba konture (engl. *rising and falling slopes*) te minimalnih i maksimalnih platoa (engl. *minima and maxima plateaux*), u slučaju klasifikacije diskretnih emocionalnih stanja [56]. Pritom je za analizu korištena DES (engl. *Danish Emotional Speech*) baza [9]. Ove su mjere korištene i za klasifikaciju govornih stilova pod stresom [57], za što su korišteni uzorci iz SUSAS (engl. *Speech Under Simulated and Actual Stress*) baze [17]. Nadalje, analiza koreliranosti parametara nagiba konture s dimenzijskim modelima emocija pokazala je negativnu korelaciju strmosti padajućih nagiba i ugode, dok su se strmosti i rastućih i padajućih nagiba pokazale kao pozitivno korelirane s pobuđenošću [43].

Složenije mjere koje su bazirane na promjenama fundamentalne frekvencije glasa, poput fluktuacije u trajanju glotalnih ciklusa (engl. *jitter*), pokazale su se kao vrlo značajne u analizi mentalnih poremećaja uzrokovanih stresom, posebice u slučaju anksioznosti [14]. Pored *jittera*, u analizi emocija i stresa koriste se i druge mjere malih varijacija (perturbacija) parametara glotalnih ciklusa, kao što su varijabilnosti maksimalne amplitude glotalnih ciklusa (engl. *shimmer*) i pravilne modulacije trajanja glotalnih ciklusa (engl. *tremor*) [27], [37], [42]. Primjerice, za šest različitih govornih stilova pod stresom iz SUSAS baze, dodavanje *jitter* mjere u bazični skup značajki rezultirao je 68.1%-tnom točnosti klasifikacije, odnosno unaprjeđenjem za nešto više od 2.5% u odnosu na korištenje samo bazičnog skupa (65.5%). Proširenje bazičnog skupa sa mjerom *shimmer* još je nešto više unaprijedilo rezultate (68.5%), dok je očekivana maksimalna točnost od 69.1% postignuta proširenjem skupa s obje perturbacijske mjere [27]. U literaturi su korištene različite matematičke metode za izračun navedenih mjera, a u [62] su one pregledno i opisane.

HARMONIJSKA STRUKTURA

Spektralna karakteristika govornog signala u osnovi sadrži informaciju o procesima u vokalnom traktu, ali i na glasnicama, s obzirom da je glas nastao konvolucijom harmonične pobude i impulsnog odziva vokalnog trakta. Brojni radovi estimiraju parametre glotalnih pulseva dekompozicijom navedenih procesa, odnosno izdvajanjem utjecaja vokalnog trakta iz govornog signala [21]. Reziduali u govornom signalu tako predstavljaju glotalne procese na grkljanu i opisani su harmonijskom strukturom u spektralnoj domeni.

Prepoznavanje emocija korištenjem parametara harmonijske strukture glasa, predloženo je u radovima [29] i [31]. Kako se takvi parametri odnose isključivo na fonacijsku informaciju iz glasa, utjecaj artikulacijskog procesa (odnosno vokalnog trakta) uklonjen je korištenjem metode predložene u [51]. Pritom se doprinos *i*-tog formanta (to jest njegove centralne frekvencije F_i i širine frekvencijskog pojasa B_i) spektru na frekvenciji f računa kao [12]:

$$V(f;F_i,B_i)=\frac{F_i^2+\left(\frac{B_i}{2}\right)^2}{\sqrt{\left((f-F_i)^2+\left(\frac{B_i}{2}\right)^2\right)\left((f+F_i)^2+\left(\frac{B_i}{2}\right)^2\right)}} \tag{1}$$

Korištenjem (1) tako su uklonjeni utjecaji prva četiri formanta na amplitude prvog i drugog harmonika H_1 i H_2, pozicioniranih na frekvencijama F_0 (fundamentalna frekvencija) i $2F_0$, kao i utjecaji formanata na amplitude spektralnih nadvišenja (engl. *peaks*) u okolici formanata A_{1p}, A_{2p} i A_{3p} na frekvencijama F_{1p}, F_{2p}, odnosno F_{3p}:

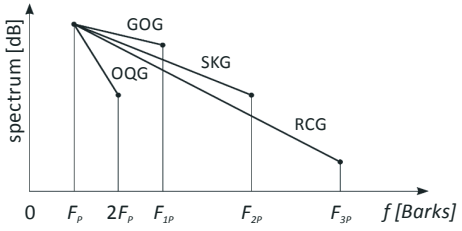
$$\begin{aligned}\tilde{H}_k &= H_k-\sum\nolimits_{i=1}^{4}V_{dB}\left(kF_0;F_i,B_i\right),\quad (k=1,2)\\ \tilde{A}_{kp} &= A_{kp}-\sum\nolimits_{\substack{i=1\\ i\neq k}}^{4}V_{dB}\left(F_{kp};F_i,B_i\right).\quad (k=1,2,3)\end{aligned} \tag{2}$$

Parametri harmonijske pobude korišteni u [29], to jest *Open Quotient Gradient* (OQG), *Glottal Opening Gradient* (GOG), *SKewness Gradient* (SKG) i *Rate of Closure Gradient* (RCG), definiraju se kao:

$$\begin{aligned}OQG &= \frac{\tilde{H}_1-\tilde{H}_2}{F_0},\\ GOG &= \frac{\tilde{H}_1-\tilde{A}_{1p}}{F_{1p}-F_0},\\ SKG &= \frac{\tilde{H}_1-\tilde{A}_{2p}}{F_{2p}-F_0},\\ RCG &= \frac{\tilde{H}_1-\tilde{A}_{3p}}{F_{3p}-F_0}.\end{aligned} \tag{3}$$

Sve frekvencijske vrijednosti pritom su pretvorene u *Bark* skalu. Ilustracija navedenih mjera prikazana je na slici 4. Kako je harmonijska struktura govora izračunata za svaki kratkotrajni vremenski otvor, tijekom izgovora se za svaki od navedenih parametara formiraju konture, kao što je prikazano na drugoj razini na slici 1. Značajke korištene za estimaciju emocija pritom se računaju na trećoj razini, odnosno uglavnom kao srednja vrijednost navedenih kontura za dani izgovor.

U [31] je OQG mjera dodatno proširena tako da obuhvaća relativne odnose prvog harmonika i narednih 20 harmonika. Te su mjere, zajedno sa skupom prozodijskih značajki, rezultirale 75.5%-tnom točnosti klasifikacije šest različitih emocija iz njemačke baze emocionalnog govora [4], odnosno doprinijele su povećanju točnosti za malo manje od 3% (72.8%) u odnosu na bazični prozodijski skup.



Slika 4: *Značajke iz harmonijske strukture glasa (preuzeto i prilagođeno iz [29]).*

Mjere kvalitete glasa, sveobuhvatno opisane u [21], koriste se i u drugim radovima za prepoznavanje emocija. Unaprjeđenje klasifikacije emocija korištenjem navedenih mjera tako je zabilježeno u radovima: [28], [52], [53] i drugim.
Pored navedenih mjera koje opisuju čisti fonacijski proces, za prepoznavanje emocija korištena je i mjera koja prikazuje omjer energija periodičkih i aperiodičkih komponenata u spektru glasa, popularnije nazvana omjer harmonika i šuma (engl. *harmonic-to-noise ratio*, HNR) [47]. »Šum« se u ovom slučaju odnosi na artikulacijsku komponentu glasa, koja je aperiodična. Pregled metoda za izračun HNR-a prikazan je u [49].

SPEKTRALNA KARAKTERISTIKA

Vremenski kratkotrajni spektar diskretnog govornog signala računa se pomoću diskretne Fourierove transformacije (engl. *discrete Fourier transform*, DFT) nad kratkotrajnim vremenskim otvorom koji čuva kvazi-stacionarna svojstva glasa te sadrži informaciju iz fonacijskog i artikulacijskog procesa. Ovako izračunati spektri nad vremenskim odsječcima kroz cijeli izgovor formiraju tzv. spektrogram. U [1] su za prepoznavanje emocija predložene značajke koje su računate iz količine energije pojaseva vremenski usrednjenih kratkotrajnih spektara, dobivenih zasebno iz zvučnih i bezvučnih dijelova izgovora.
Tako su iz vremenski usrednjenog spektra nad zvučnim dijelovima izgovora značajke računate kao količine energije sljedećih spektralnih pojaseva: 125 – 200 Hz, 200 – 300 Hz, 300 – 500 Hz, 500 – 600 Hz, 600 – 800 Hz, 800 – 1000 Hz, 1000 – 1600 Hz, 1600 – 5000 Hz i 5000 – 8000 Hz. U analizu je uključena i razlika maksimalnih energija pojasa 0 – 2000 Hz u odnosu na pojas 2000 – 5000 Hz. Od dodatnih mjera usrednjenog spektra zvučnog dijela izgovora korišten je spektralni gradijent (tzv. *drop-off*), to jest nagib aproksimativnog pravca na frekvencijskom rasponu od 1000 Hz na više, te relativna količina energije viših u odnosu na niže pojaseve, uz pragove eksperimentalno postavljene na 500 Hz i 1000 Hz.
Iz usrednjenog spektra nad bezvučnim dijelom izgovora značajke su računate kao količine energija nad pojasevima: 125 – 250 Hz, 250 – 400 Hz, 400 – 500 Hz, 500 – 1000 Hz, 1000 – 1600 Hz, 1600 – 2500 Hz, 2500 – 4000 Hz, 4000 – 5000 Hz i 5000 – 8000 Hz. U [47] je za prepoznavanje emocija korišten spektralni *flux*, dobiven iz spektrograma, dok su neke značajke, poput centra gravitacije spektra i tzv. *roll-off* točke, izračunate iz spektralne karakteristike cjelokupnog govornog signala.
Postupkom usrednjavanja spektrograma kroz izgovor gubi se informacija o fonacijskom i artikulacijskom procesu i analiziraju se samo informacije o zastupljenosti pojedinih spektralnih komponenti kroz izgovor, to jest izraženosti energija pojedinih spektralnih pojaseva. Izvlačenje informacija o fonacijskom procesu opisano je ranije u ovom poglavlju, dok je informacije o artikulacijskom procesu moguće promatrati kroz analizu MFCC koeficijenata (engl. *Mel-scale Frequency Cepstral Coefficients*) [32], LPC koeficijenata (engl. *Linear Prediction Coefficients*) [39] i drugih sličnih mjera koje opisuju oblik i položaj vokalnog trakta prilikom izgovora. Analiza MFCC koeficijenata, isto kao i kod prepoznavanja govora, uglavnom obuhvaća i dinamičku informaciju u obliku delta i akceleracijskih koeficijenata [47], dok se u [31] koriste centralni momenti izračunati nad

konturama 13 MFC koeficijenata tijekom izgovora. MFCC mjere nisu pokazale značajniju korelaciju s emocijama [35], no često se koriste u sklopu bazičnog skupa, s kojim se onda uspoređuju novo-predloženi skupovi značajki [31].
Za analizu promjena artikulacijskog procesa uslijed djelovanja emocija često se koriste formanti, odnosno karakteristična nadvišenja u kratkotrajnoj spektralnoj karakteristici glasa, koja se uglavnom računaju na temelju LPC analize. Ovdje se pored centralne frekvencije formanata prate i promjene širine njihovog frekvencijskog pojasa (engl. *bandwidth*) tijekom izgovora. Na ovaj način se u *z*-ravnini vrši analiza kuta ali i radijusa konjugirano kompleksnih parova polova prijenosne funkcije vokalnog trakta. Utjecaj emocija na promjene formanata detaljno je prikazan u [41], a u velikoj se mjeri temelji na istraživanju iz [23].

OSTALE MJERE

Kontura intenziteta (energije ili snage) govornog signala izračunata pomicanjem kratkotrajnog vremenskog otvora duž signala, uz fundamentalnu frekvenciju i trajanje izgovora čini prozodijsku strukturu govora, koja se smatra osnovicom za prepoznavanje emocija [31]. Osnovne akustičke značajke iz konture intenziteta (kao što su srednja vrijednost, raspon i standardna devijacija) predložene su u [41], no u radovima se (slično kao i za fundamentalnu frekvenciju) često koristi cijeli niz statističkih mjera za opis ove konture [31], [56].
U [56] i [57] se značajke dodatno računaju i kao statističke mjere iz izdvojenih rastućih i padajućih nagiba konture, te nad minimalnim i maksimalnim platoima konture. Skup uobičajenih statističkih značajki dodatno je proširen u [45], gdje se iz konture kratkotrajne energije računa i relativni iznos maksimuma konture u odnosu na srednju vrijednost tijekom izgovora te vremenski trenutak u izgovoru u kojem je maksimum ostvaren. Računaju se dodatno i srednja vrijednost i standardna devijacija udaljenosti između infleksijskih točaka konture energije tijekom izgovora.
Značajke iz konture intenziteta su, pored dobre korelacije s diskretnim emocionalnim stanjima, pokazale značajnu korelaciju i s dimenzijskim modelima emocija. Tako je u [43] ustanovljena pozitivna korelacija medijana i raspona intenziteta kroz izgovor s ugodom i pobuđenošću.
Brzina izgovora (engl. *speech rate*) definirana je brojem izgovorenih riječi u nekoj jedinici vremena [41] i u literaturi je, u kontekstu prepoznavanja emocija, objedinjena krovnim terminom ‘trajanje’ (engl. *duration*) [31]. U radovima se, pored brzine izgovora, analizira trajanje riječi kroz izgovor, kao i trajanje zasebno zvučnih i bezvučnih dijelova, te intervala tišine [45]. U [24] je naglasak stavljen na analizu međusobnih omjera trajanja zvučnih, bezvučnih i intervala tišine. Ovako definirane značajke ne podliježu klasičnom izračunu pomoću kratkotrajnih vremenskih otvora, opisanom ranije u ovom poglavlju.
Brzina prolazaka govornog signala kroz nulu (engl. *zero-crossing rate*, ZCR) vrlo je česta i pouzdana mjera za prepoznavanje emocija. Prema nekim autorima (npr. [38]) ZCR se smatra jednom od mjera iz kategorije trajanja izgovora, dok se prema drugima ona gleda kao zasebna kategorija značajki (npr. [31]). Njezina prednost u odnosu na brzinu

izgovora je u tome što je nevezana uz jezik [38]. Osnovne statističke mjere poput srednje vrijednosti, minimalne i maksimalne vrijednosti te sličnih mogu se primijeniti i ovdje ukoliko se ZCR računa nad kratkotrajnim vremenskim otvorima kroz izgovor.

U dosad navedenim mjerama obuhvaćene su gotovo sve akustičke značajke iz glasa za prepoznavanje emocija, korištene u literaturi. Inovativan pristup izračunu značajki predložen je u [20], gdje je analizirana razlika između klasičnih značajki dobivenih na temelju kratkotrajnih vremenskih otvora (engl. *frame-level*) te značajki dobivenih na razini cijelog izgovora (engl. *utterance-level*). Značajke na razini izgovora računate su metodom dekompozicije *wavelet* paketa u 5 razina. Skup značajki dobiven je tako iz iznosa energija svih 32 frekvencijskih opsega. Fuzijom oba skupa značajki postiže se povećanje točnosti klasifikacije i do 13.6% u odnosu na klasifikaciju emocija samo pomoću *frame-level* skupa značajki. U [45] su pojedine značajke također računate iz cjelokupnog izgovora i to direktno iz govornog signala. Određene statističke mjere, poput srednje vrijednosti i medijana »sirovog« govornog signala (engl. *raw speech*), pokazale su se kao značajne za prepoznavanje sedam diskretnih emocija (6 primarnih i neutralno stanje).

PRIMJENE METODA STROJNOG UČENJA

U literaturi se za klasifikaciju diskretnih emocionalnih stanja najčešće primjenjuju modeli s Gaussovim mješavinama (engl. *Gaussian mixture models*, GMM). U [45] se GMM-ovi koriste za klasifikaciju sedam emocionalnih stanja, od čega šest stanja pripada primarnim emocijama, a jedno se odnosi na neutralno stanje. Za sustav neovisan o govorniku dobiva se prosječna pogreška od 25.17%, dok se kod sustava ovisnog o govorniku ona smanjuje na 10.88%. Redukcija prostora akustičkih značajki u tom radu na 33 dimenzije je provedena korištenjem LDA postupka. Rezultati ove metode su uspoređeni s rezultatima ostalih metoda za isti skup uzoraka i isti broj diskretnih emocija, što je prikazano u tablici 1. U [33] se koriste odvojeni skupovi Gaussovih komponenata, i to njih 512 za MFCC značajke ($M_1 = 512$), a 64 komponente za značajke iz *pitcha* ($M_2 = 64$). Za finalnu adaptaciju tijekom učenja modela EM algoritmom korišten je MAP kriterij re-estimacije parametara. Klasifikacija se vršila nad tri klase emocija: neutralne, empatičke i negativne (frustracije). Postignute točnosti dosežu i do 95% u najboljem slučaju.

U [31] su za analizu unaprjeđenja točnosti klasifikacije uz primjenu značajki kvalitete glasa također korišteni GMM-ovi, s time da je selekcija skupa od 25 relevantnih značajki provedena pomoću SFFS metode. U slučaju proširenog skupa značajki (sastavljenog od prozodijskih značajki i značajki kvalitete glasa) postignuta je točnost klasifikacije od 75.5%, dok su u slučaju hijerarhijske 2-stupanjske i 3-stupanjske klasifikacije postignute točnosti od 83.5%, odnosno 88.8% za šest diskretnih emocija. Inovativno rješenje za treniranje GMM modela realizirano je u radu [60]. Ovdje se problem pretreniranosti (engl. *over-fitting*) rješava pomoću rubnog skaliranja (engl. *margin scaling*) u svrhu generalizacije modela u procesu učenja. Takvo skaliranje provodi se pomoću funkcija gubitaka (engl. *loss functions*), koje definiraju udaljenost pojedinih klasa emocija na temelju grupacije *Watson-Tellegen* modelom [58]. Eksperimentalnom evaluacijom je utvrđeno

da ovakvo učenje rezultira većom točnosti klasifikacije emocija u odnosu na ostale, klasične metode učenja GMM-ova.

Uz standardnu primjenu skrivenih Markovljevih modela (engl. *hidden Markov models*, HMM) za prepoznavanje govora (prepoznavanje teksta iz audio zapisa govora), vrlo je česta njihova primjena i za prepoznavanje emocija. U [34] se HMM-ovi s 64 stanja koriste za klasifikaciju 7 diskretnih emocija (6 primarnih i neutralno stanje), gdje se postiže dosta visoka točnost od 82.5%. Slične točnosti, s nešto drugačijim akustičkim značajkama, postigao je Schuller u radu [44]. Dodatno je analiziran i utjecaj broja stanja svakog modela na točnost prepoznavanja. Pokazalo se da se ukupna točnost prepoznavanja povećava od 65% za modele s jednim stanjem (što odgovara GMM modelu) do preko 80% za 64 stanja svakog pojedinog modela. U [35] je pokazan drugačiji odnos točnosti i broja stanja HMM-ova, gdje je najveća točnost postignuta za modele s 4 stanja. Klasifikacija je rađena nad istim skupom emocija i nešto drugačijim skupom akustičkih značajki. Zanimljiv koncept je prikazan u [30], gdje je za prepoznavanje četiri tipa kvalitete glasa prema Laveru [23]: modalni, zadihan, grub i škripav, korišten samo jedan HMM s četiri stanja. Svako stanje na ovaj način predstavlja jedan tip glasa te se tijekom izgovora estimira najvjerojatniji prolaz kroz stanja pomoću Viterbijevog algoritma, odnosno najvjerojatniji tijek promjene kvalitete glasa kroz izgovor. U tu svrhu su opservacijski vektori definirani na temelju šest kontura, odnosno mjera kvalitete glasa. Ovdje je koncept HMM-a iskorišten u punom smislu te se klasifikacija ne primjenjuje nad cjelokupnim izgovorom (kao u ostalim radovima), nego se odvija unutar izgovora. Izgovor je na ovaj način moguće podijeliti na segmente u ovisnosti o kvaliteti glasa tijekom izgovora, što je značajna funkcionalnost s obzirom na dinamična i vremenski promjenjiva svojstva glasa. Klasifikacija primarnih diskretnih emocija korištenjem strojeva s potpornim vektorima (engl. *support vector machines*, SVM) je prikazana u radu [45]. Za klasifikaciju su korištene akustičke glasovne značajke uz transformaciju prostora pomoću LDA metode. Isprobane su tri arhitekture: klasičan pristup *jedan-protiv-svih*; kombinacija SVM-a s umjetnom neuronskom mrežom (MLP); te višeslojna struktura SVM-ova, gdje se kroz četiri sloja binarnim klasifikatorima dolazi do finalnih sedam klasa diskretnih emocija (ML-SVM). Najbolji rezultat za sustav neovisan o govorniku ostvaruje se višeslojnom arhitekturom SVM-ova (18.71% pogreške), dok klasičan SVM postiže najmanju pogrešku za sustav ovisan o govorniku (7.05%). Ovaj rezultat, zajedno s rezultatima ostalih metoda, prikazan je u tablici 1. Za mapiranje je korištena Gaussova RBF funkcija. SVM-ovi su korišteni i kod klasifikacije kategorija dimenzijskih emocija (ugode i pobuđenosti) u [59]. Tu je 2-dimenzionalna ravnina podijeljena na 16 klasa u jednom slučaju (4 za svaku dimenziju) te 49 klasa u drugom (7 za svaku dimenziju). Rezultati za pobuđenost dosežu točnost od 46.3% za 16 klasa te 30.8% za 49 klasa. Za ugodu je točnost prepoznavanja još manja i iznosi 44.6% u prvom te 24.6% u drugom slučaju. Rađene su i analize točnosti sustava za klasifikaciju s mogućnošću konfuzije među graničnim klasama, što je znatno popravilo rezultate.

Visoke točnosti klasifikacije emocija korištenjem SVM-a su postignute u [19]. Ovdje je jezgra (engl. *kernel*) SVM-a formirana na temelju GMM-a, što je rezultiralo dosta viso-

kom točnosti klasifikacije pet diskretnih emocija: 82.5% za miješanu populaciju govornika, 91.4% u slučaju muške populacije i 93.6% u slučaju ženske. Za usporedbu, isti zadatak realiziran korištenjem klasičnog GMM klasifikatora je rezultirao nešto lošijim točnostima: 77.9% za miješanu populaciju, 85.7% za mušku i 87.3% za žensku. Estimacija dimenzija emocija: ugode, pobuđenosti i dominacije, realizirana je u [16] korištenjem SVR-a (engl. *support vector regression*) uz isprobanu linearnu, polinomnu i RBF jezgru. Optimalni parametri su pronađeni zasebno za svaku jezgru. Najbolji rezultati u formatu srednje pogreške (engl. *mean error*, ME) su za sve tri dimenzije emocija postignuti u slučaju korištenja RBF jezgre. Tako su u slučaju ugode, pobuđenosti i dominacije postignute srednje pogreške u iznosu od 0.13, 0.15, odnosno 0.14, uz raspon dimenzija [-1, 1]. Binarno stablo odluke korišteno je za klasifikaciju diskretnih emocija u radu [25]. Realizirana je hijerarhijska arhitektura s nekoliko razina binarne klasifikacije s ciljem da se prvo provedu najlakši klasifikacijski zadaci. U čvorovima grananja se u jednom slučaju primjenjivao SVM, a u drugom Bayesova logistička regresija (engl. *Bayesian Logistic Regression*). Obje metode su ostvarile zadovoljavajuće rezultate. U [45] su za klasifikaciju sedam diskretnih emocionalnih stanja iz akustičkih značajki, pored već navedenih metoda, isprobane i kNN (metoda *k* najbližih susjeda), kMeans (metoda *k* srednjih vrijednosti) te MLP. Nijedna od ovih metoda nije nadmašila rezultate SVM metoda, što je vidljivo u tablici 1.

Tablica 1: *Pogreške klasifikacije 7 diskretnih emocija na temelju akustičkih značajki (preuzeto i prilagođeno iz [45]).*

Klasifikator	Sustav neovisan o govorniku	Sustav ovisan o govorniku
kMeans	57.05 %	27.38 %
kNN	30.41 %	17.39 %
GMM	25.17 %	10.88 %
MLP	26.85 %	9.36 %
SVM	23.88 %	7.05 %
SVM-MLP	24.55 %	11.3 %
ML-SVM	18.71 %	9.05 %

U [59] je provedena usporedna analiza više metoda strojnog učenja na problemu estimacije dimenzija ugode i pobuđenosti iz glasovnih značajki. Za klasifikaciju kategorija dimenzijskih emocija (16 u jednom primjeru i 49 u drugom) korišteni su SVM-ovi i uvjetna slučajna polja (engl. *conditional random fields*, CRF). Točnost rezultata klasifikacije korištenjem CRF-ova je veća od točnosti korištenjem SVM-ova (koja je navedena ranije). U slučaju pobuđenosti CRF klasifikator tako postiže točnost od 50.8% za 16 klasa te 32.5% za 49 klasa, dok je u slučaju ugode postignuta točnost prepoznavanja od 45.6% u prvom te 29.7% u drugom slučaju. Usporedna analiza je provedena i na problemu estimacije kontinuiranih vrijednosti. U tu svrhu su korišteni SVR-ovi i LSTM-RNN metoda (engl. *long short-term memory recurrent neural network*). U slučaju pobuđenosti LSTM-RNN metoda rezultira manjom srednjom kvadratnom pogreškom (engl. *mean square error*,

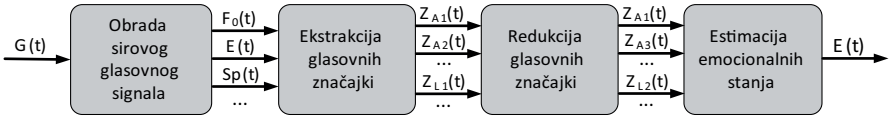
MSE) u odnosu na SVR-ove (LSTM-RNN: 0.08, a SVR: 0.10), dok je MSE u slučaju ugode jednak za najbolje slučajeve obje metode i iznosi 0.18. Iznosi ugode i pobuđenosti su definirani u rasponu [-1, 1].

LSTM-RNN struktura opisana u [10] i [59] je u potpunosti prilagođena za estimaciju dimenzija emocija (ugode i pobuđenosti) u stvarnom vremenu. U tu svrhu su analizirane i akustičke i lingvističke značajke. Ideja je bila iskoristiti memorijsko svojstvo takve strukture te na ulaz svakih 20 ms dovoditi novi skup značajki i na taj način estimirati emocije u stvarnom vremenu. Akustičke se značajke računaju pomoću vremenskih otvora širine 32 ms, s pomakom od 20 ms. Ukupan skup od 56 značajki dobiven je od 28 akustičkih parametara, zajedno s isto toliko delta koeficijenata. Ovakvo dinamično računanje značajki, koje se u radu nazivaju zvučnim deskriptorima niske razine (engl. *low-level audio descriptors*, LLD) i određuju se neovisno o susjednim uzorcima u izgovoru, omogućuje brzu i kontinuiranu informaciju na izlazu. Analiza lingvističkih značajki, odnosno pojavljivanja 56 ključnih riječi u izgovoru (definiranih vokabularom) je omogućena korištenjem memorijskog spremnika (engl. *buffer*). Za obuhvaćanje dugotrajnih (engl. *long-range*) svojstava izgovora, nužnih za potpunu informaciju o emocionalnom stanju, korištene su LSTM ćelije. Rezultati su u [10] prikazani kroz mjere korelacije rezultata i referentnog znanja. Takva je metoda prikaza, prema tvrdnji autora, robusnija na utjecaje pomaka (engl. *offset*) i šuma tijekom estimacije od linearnih i kvadratnih mjera pogrešaka. Postižu se iznosi korelacije nešto manji od 0.5 za pobuđenost te isto tako i za ugodu.

SUSTAV ZA ESTIMACIJU EMOCIONALNIH STANJA NA TEMELJU DUBINSKE ANALIZE AKUSTIČKIH ZNAČAJKI GOVORA

Arhitektura sustava za estimaciju emocionalnih stanja na temelju akustičkih značajki govornog signala je u generičkom formatu prikazana na slici 5 i sadrži sljedeće komponente:

- *obrada govornog signala* – Izvlačenje akustičkih mjera poput F_0, energije, formanata i spektralne karakteristike iz govornog signala u digitalnom formatu.
- *izračun akustičkih značajki* – Formiranje skupa akustičkih značajki iz gore navedenih akustičkih mjera govora.
- *redukcija značajki* – Selekcija relevantnih značajki iz početnog skupa i/ili transformacija prostora značajki.
- *estimacija emocionalnih stanja* – Klasifikacija ili regresija opservacijskih vektora značajki na kategorije, odnosno dimenzijske vrijednosti emocionalnih stanja pomoću metoda strojnog učenja.



Slika 5: *Generička arhitektura sustava za estimaciju emocionalnih stanja na temelju akustičkih značajki govornog signala.*

Iz prikupljenih se uzoraka govornog signala računaju razne akustičke mjere, odnosno konture, a potom i značajke koje formiraju početni skup. Takav je skup značajki često velik i redundantan, te se podvrgava nekoj od metoda redukcije značajki. Metoda redukcije značajki podrazumijeva selekciju relevantnih značajki i/ili transformaciju prostora značajki, što rezultira manjim i robusnijim prostorom značajki. Postupak reduciranja prostora značajki, odnosno smanjenja dimenzionalnosti opservacijskog vektora značajki, primjenjuje se uglavnom radi prilagodbe sustava na količinu trening uzoraka. U posljednjoj fazi se na temelju opservacijskog vektora značajki estimira emocionalno stanje nekom od metoda strojnog učenja.

Metoda strojnog učenja za estimaciju emocionalnih stanja i parametri njezinih modela odabiru se uglavnom prema ciljanoj reprezentaciji emocija, te prema dimenzionalnosti opservacijskog vektora značajki i količini uzoraka za trening modela. Metode klasifikacije se tako koriste u slučaju diskretnih emocija, poput sreće, tuge, straha, ljutnje i sličnih, dok se regresijske metode koriste za estimaciju dimenzijskih veličina emocija, poput ugode i pobuđenosti, te intenziteta diskretnih emocija i stresa.

U razvojnoj fazi sustava za estimaciju emocionalnih stanja iz akustičkih značajki potrebno je razraditi sljedeće stavke:

- ovisnost sustava o subjektu (personalizacija);
- skup glasovnih značajki (akustičke / lingvističke / fuzija);
- tip estimacije emocija (klasifikacija / regresija);
- metoda za estimaciju emocionalnih stanja;
- vremenski tijek estimacije emocija (stvarno vrijeme / bez ograničenja stvarnog vremena).

Pritom je važno osigurati potrebnu količinu i kvalitetu uzoraka za treniranje te validaciju modela sustava. U radovima se uglavnom koriste već formirane i evaluirane baze emocionalnog govora, no ponekad se zbog specifične problematike istraživanja formiraju nove baze. Takav proces obuhvaća planiranje i provođenje eksperimenta tijekom kojeg se prikupljaju govorni uzorci, koji se potom anotiraju odgovarajućim emocionalnim stanjima, najčešće od strane nezavisnih slušača. Kod planiranja eksperimenta, odnosno definiranja eksperimentalnog protokola, potrebno je definirati sljedeće stavke:

- ciljana skupina ispitanika (najčešće neka specifična skupina koja se uspoređuje s kontrolnom skupinom ispitanika);
- oblik ekspresije emocija (realističan (naturalističan / induciran pobudama) ili glumljen);
- oblik pobuda (govorna ili tekstualna pitanja, statične slike, zvukovi, video materijali ili virtualna stvarnost);
- strategija pobuđivanja (definiranje sekvenci, seansi i dolazaka);
- ciljano ponašanje ispitanika (fonacija / izgovaranje unaprijed naučenog teksta / čitanje zadanog teksta / verbalni kognitivni zadatci (matematičke operacije) / komentiranje pobude / odgovaranje na zadana pitanja / spontani govor / dijalog);
- način reprezentacije emocija (diskretan i/ili dimenzionalan);
- način anotacije emocija na govornim uzorcima – referentno znanje (subjektivni iskazi, mišljenje eksperata i slično).

Kategorizacija eksperimenta prema obliku ekspresije emocija podijeljena je na realistične i glumljene emocije. Realistične emocije se dodatno dijele na naturalistične, koje su nastale bez unaprijed planiranog scenarija (na primjer tijekom dijaloga na info pultovima ili nekih situacija iz stvarnog života), i na ekspresije koje su inducirane pobudama u nadziranom okruženju (na primjer tijekom terapija izlaganjem ili upravljanja simulatorom). Na području estimacije emocija iz govora se najčešće rade eksperimenti u kojima se prikupljaju realistične emocije. Djelomičan uzrok tome je u sociološkim i biološkim razlozima uslijed kojih je potreban određen intenzitet emocionalne reakcije da se uoči »prava« promjena u glasu, a kojeg je lakše ostvariti u naturalističnom okruženju [6]. Scherer tvrdi da takav intenzitet nije nužno vezan uz naturalističko okruženje s obzirom da: »... su naturalistične ekspresije djelomično glumljene, kao što su i glumljene ekspresije djelomično naturalistične« [1].

Kod eksperimenata gdje se emocije induciraju pobudama koriste se različiti multimedijski formati. Baza slika IAPS (engl. *International Affective Picture System*) [22] i baza zvukova IADS (engl. *International Affective Digitized Sounds*) [3] često su korištene u takve svrhe. Te su baze prikupljene i anotirane dimenzijskim vrijednostima ugode, pobuđenosti i kontrole na Sveučilištu u Floridi.

Kod razrade strategije pobuđivanja, odnosno definiranja eksperimentalnog protokola, potrebno je definirati broj dolazaka ispitanika i termine te odrediti strategiju pobuđivanja tijekom seansi. Seansa se sastoji od niza multimedijskih pobuda različitog emocionalnog sadržaja i semantike tijekom kojih se prikupljaju govorni uzorci, to jest reakcije. Svaka seansa može imati nekoliko sekvenci, odnosno tematski povezanih podnizova multimedijskih pobuda.

U završnoj fazi eksperimenta se vrši anotacija govornih ekspresija emocija odgovarajućim emocionalnim oznakama. Postoji nekoliko načina takvog označavanja, kao što su: subjektivni iskazi, korištenje postojeće anotacije pobude pri označavanju ekspresije, procjene eksperata, primjene znanja iz literature i slično.

Prilikom formiranja sustava za estimaciju emocionalnih stanja iz glasa treba biti svjestan sljedećih ograničenja:

- Većina sustava za estimaciju emocija funkcionira na principu prethodno utreniranih modela emocija, te je od velike važnosti prikupiti uzorke za gradnju modela koji po svojoj karakteristici, što je moguće bliže, odgovaraju uzorcima iz stvarnog okruženja u kojem će se sustav za estimaciju koristiti.
- Kako emocionalne reakcije u glasu mogu interferirati i s ostalim afektivnim stanjima (npr. raspoloženjem), te s drugim parametrima koji također utječu na glas (npr. umor, iscrpljenost, bolest, unos kave i ostalih namirnica i slično), trebalo bi takve interakcije uzeti u obzir prilikom procesa anotacije emocionalnog govora.
- Parametri modela sustava za estimaciju emocionalnih stanja se postavljaju na temelju anotacija, odnosno referentnog znanja o emocijama zastupljenim u govornim uzorcima. Referentno znanje se temelji na nekoj od gore opisanih metoda anotacije. Ukoliko se, primjerice, anotacija temelji na subjektivnom iskazu ispitanika, onda treba biti svjestan da se pod pojmom estimacija emocionalnih stan-

ja misli na prepoznavanje tog subjektivnog osjećaja ispitanika, odnosno njegovog doživljaja vlastite emocionalne reakcije. To je prema Schereru samo jedna komponenta emocije, koja nije nužno u korelaciji s ostalim komponentama emocije (poput fiziološke i motoričke reakcije) koje se uglavnom mjere iz glasovnih značajki [15].

BAZE EMOCIONALNOG GOVORA

Baze emocionalnog govora se u osnovi dijele prema obliku ekspresije emocija. Najpoznatije baze su navedene u nastavku. Često citirana baza glumljenih emocija formirana je za potrebe rada [1]. Dvanaest profesionalnih glumaca (6 muškaraca i 6 žena), izvornih njemačkih govornika, pristupilo je zadatku. Izgovarali su dvije rečenice, komponirane od fonema različitih indoeuropskih jezika, bez semantičkog značenja:

„*Hat sundig pron you venzy*»,
„*Fee gott laish jonkill gosterr*».

Kako ne postoji lingvistička informacija, akustička komponenta je maksimalno izražena. Govornici su izgovarali rečenice za svako od 14 diskretnih emocionalnih stanja u nekoliko varijanti, što je rezultiralo bazom od ukupno 1344 izgovora.

HUMAINE je također često korištena i citirana baza emocionalnog govora (realističnih ekspresija emocija), koja je formirana u sklopu projekta na Sveučilištu u Belfastu [8]. Ovo je po obujmu najveća baza emocionalnog govora, sastavljena od nekoliko postojećih baza: *Belfast Naturalistic Database*, *EmoTV Database*, *Castaway Reality Television Database*, *Sensitive Artificial Listener*, *Activity Data/Spaghetti Data*, *Belfast Driving simulator Data*, *EmoTABOO*, *Green Persuasive Dataset* te *DRIVAWORK*. Prve tri baze su formirane od naturalističnih ekspresija emocija, a ostale od ekspresija induciranih pobudama.

Baza govora u uvjetima simuliranog i stvarnog stresa (engl. *Speech Under Simulated and Actual Stress*, SUSAS) nastala je na sveučilištu Colorado-Boulder [17]. Standardizirana je na način da su govornici u raznim uvjetima izgovarali 35 uobičajenih riječi iz komunikacije zračnog prometa. Sadrži pet domena govora pod stresom [18]:

1. *govorni stilovi* – Formirana je baza sedam govornih stilova u uvjetima stresa: spor, brz, ljutit, glasan, mekan, čist i upitan. Ovdje se radi o govoru u uvjetima simuliranog stresa, gdje je ukupno devet muških govornika 8820 puta izgovorilo standardizirane riječi.
2. *zadatak s jednostrukim praćenjem* – Uzorci su snimani za vrijeme simuliranih uvjeta stresa (umjerenog i povišenog) tijekom izvršavanja računalno zadanih zadataka, odnosno kalibriranog radnog opterećenja (engl. *workload*). Ovdje je istih devet govornika izgovaralo ukupno 1890 puta riječi.
3. *zadatak s dvostrukim praćenjem* – Ovdje su simulirani uvjeti kontrole leta i praćenje cilja, također u uvjetima umjerenog i izraženog stresa. Osam govornika (četiri muškarca i četiri žene) su izgovorila ukupno 2257 puta standardizirane riječi.
4. *govor pod stvarnim stresom* – Ova domena se sastoji od dva skupa (iz kojih je ukupno 11 govornika, osam muškaraca i tri žene, izgovorilo 1642 puta standardizirane riječi):

- snimki leta Apache helikopterom, gdje su dva muška pilota izgovarala standardizirane riječi na tlu i u uvjetima leta (dodatno postoje i snimke govora dva muška pilota tijekom leta u uvjetima nestanka goriva – snimke čistog govora);
- snimki iz zabavnog parka i to tijekom slobodnog pada i tijekom vožnje tzv. vlakom smrti (engl. *roller-coaster*).

5. *psihoterapija* –Ukupno osam pacijenata (dva muškarca i šest žena) je snimano tijekom terapije u Emory medicinskom centru, Odjelu za psihijatriju. Ova domena nije dio standardne SUSAS baze. Dostupna je na zahtjev, uz posebne dozvole.

ZAKLJUČAK

Kako je glasovna komponenta ekspresije emocija još nedovoljno istražena, estimacija emocija na temelju glasovnih značajki predstavlja izazovan, i još uvijek aktivan problem u svijetu znanosti. Većina istraživanja i dalje je fokusirana na analizu diskretnih emocija, što ostavlja prostor za proučavanje glasovnih značajki u kontekstu dimenzijskih emocija (ugode i pobuđenosti). U zadnje vrijeme se sve više pažnje posvećuje mogućnosti kontinuirane estimacije emocija. Za potrebe HCI (engl. *human-computer interaction*) sustava nastoji se razviti modele koji će biti u stanju vršiti takve estimacije u stvarnom vremenu. Isto tako, nastoji se obuhvatiti što je moguće više informacija iz glasa, što rezultira konstantnom inovativnošću na području odabira značajki.

Ovo je područje, osim znanstvenim krugovima, interesantno i uslužnom sektoru zbog mogućnosti unaprjeđenja medicine, psihologije, obrazovanja, te nekih drugih djelatnosti u kojima je nužno što sofisticiranije automatizirati uslugu – primjerice u pozivnim centrima, automobilskoj industriji i slično. Unatoč tome, još ne postoji komercijalan i robustan sustav koji je u stanju postići zadovoljavajuću razinu točnosti prepoznavanja emocija, neovisno o subjektu i ambijentu. To je djelomično tako i zbog izuzetne kompleksnosti fenomena emocija, koje ni sam čovjek koji ih proživljava ponekad nije u stanju precizno opisati. Ovo ostavlja još puno prostora za nadogradnju, s naglaskom na multidisciplinarnu povezanost struka.

Također treba naglasiti da se estimacija emocionalnih stanja na temelju glasa može kombinirati i s ostalim metodama, npr. estimacijom na temelju izraza lica i fizioloških reakcija, kako bi se na što učinkovitiji način obuhvatila prednost i mogućnosti svake pojedine metode.

REFERENCE

[1] Banse, R., Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression, *J. Personality Social Psych*., Vol. 70, No 3, str. 614-636.

[2] Bezooijen, R., Goverina, A. M. (1984). *Characteristics and recognizability of vocal expressions of emotion.* Diss. Foris Publications.

[3] Bradley, M. M., Lang, P. J. (2007). *The International Affective Digitized Sounds*, 2nd edition, (IADS-2): Affective ratings of sounds and instruction manual, Technical report B-3, University of Florida, Gainesville, FL.

[4] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B. (2005). A Database of German Emotional Speech, *Proceedings of the Interspeech 2005*, ISCA, Lisbon, Portugal, str. 1517-1520.

[5] Cowan, M. (1936). Pitch and intensity characteristics of stage speech. *Arch. Speech.*

[6] Cowie, R. et al. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, Vol. 18, No. 1, str. 32-80.

[7] Damasio, A. R. (1999). *The feeling of what happens: body and emotion in the making of consciousness*, Harcourt Brace & Company, New York.

[8] Douglas-Cowie, E. et al. (2005). Multimodal databases of everyday emotion: Facing up to complexity. *Proceedings of the Interspeech'05*, str. 813-816.

[9] Engberg, I. S., Hansen, A. V. (1996). Documentation of the Danish Emotional Speech Database (DES), *Internal AAU report*, Center for Person Kommunikation, Denmark.

[10] Eyben, F., Wollmer, M., Graves, A., Schuller, B., Douglas-Cowie, E., Cowie, R. (2010). On-line Emotion Recognition in a 3-D Activation-Valence-Time Continuum using Acoustic and Linguistic Cues. *Journal on Multimodal User Interfaces*.

[11] Fairbanks, G., Pronovost, W. (1939). An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech monograph*, Vol. 6, str. 87-104.

[12] Fant, G. (1970). *The Acoustic Theory of Speech Production: 2nd edition*. Mouton, Hague.

[13] Frick, R. (1985). Communicating emotion: The role of prosodic features. *Psychological Bulletin*, Vol. 97, str. 412-429.

[14] Fuller, B. F., Horii, Y., Conner, D. A. (1992). Validity and reliability of nonverbal voice measures as indicators of stressor-provoked anxiety. *Research in Nurse & Health*, Vol. 15, No. 5, str. 379-389.

[15] Grandjean, D., Sander, D., Scherer, K. R. (2008). Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Consciousness and cognition*, Vol. 17, No. 2, str. 484-495.

[16] Grimm, M., Kroschel, K., Narayanan, S. (2007). Support vector regression for automatic recognition of spontaneous emotions in speech. *Proceedings of the ICAASP '07.*

[17] Hansen, J. H. L. (1996). Analysis and compensation of speech under stress and noise for enviromental robustness in speech recognition. *Speech Communication*, Vol. 20, str. 151-173.

[18] Hansen, J. H. L. et al. (2000). The impact of speech under 'stress' on military speech technology. *Nato Project 4*.

[19] Hu, H. et al. (2007). GMM supervector based SVM with spectral features for speech emotion recognition. *Proceedings of the ICAASP '07.*

[20] Huang, Y., Zhang, G., Li, X., Da, F. (2011). Improved Emotion Recognition with Novel Global Utterance-level Features. *Applied Mathematics & Information Sciences*, Vol. 5, No. 2, str. 147-153.

[21] Kane, J. (2012). *Tools for analysing the voice*, doktorski rad, Trinity College Dublin, Dublin, Ireland.

[22] Lang, P. J., Bradley, M. M., Cuthbert, B. N. (2005). International Affective Picture System (IAPS): Affective ratings of pictures and instruction manual, *Technical Report A-6*, University of Florida, Gainesville, FL.

[23] Laver, J. (1980). *The Phonetic Description of Voice Quality*, Cambridge University Press.

[24] Lee, C. M., Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Trans. Speech and Audio Processing*, Vol. 13, No. 2, str. 293-303.

[25] Lee, C-C., Mower, E., Busso, C., Lee, S., Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Elsevier Speech Communications Journal*, Vol. 53, str. 1162-1171.

[26] Lewis, M. D. (2005). Bridging emotion theory and neurobiology through dynamic systems modeling. *Behavioral and Brain Sciences*, Vol. 28, str. 169-245.

[27] Li, X. et al. (2007). Stress and emotion classification using jitter and shimmer features. *ICASSP '07*, 4, str. 1801-1804.

[28] Lliev, A. I., Scordilis, M., Papa, J., Falcao, A. (2010). Spoken emotion recognition through optimum-path forest classi cation using glottal features. *Computer Speech and Language*, Vol. 24, No. 3, str. 445-460.

[29] Lugger, M., Yang, B. (2006). Classification of different speaking groups by means of voice quality parameters. *ITG-Fachbericht-Sprachkommunikation*.

[30] Lugger, M., Stimm, F., Yang, B. (2008). Extracting voice quality contours using discrete hidden Markov models. *Proceedings of the Speech Prosody*.

[31] Lugger, M., Yang, B. (2008). Psychological motivated multi-stage emotion classification exploiting voice quality features. *Speech Recognition*, In-Tech.

[32] Merlmestein, P. (1976). Distance Measures for Speech Recognition – Psychological and Instrumental. *Joint Workshop on Pattern Recognition and Artificial Intelligence*, str. 91-103.

[33] Neiberg, D., Elenius, K., Laskowski, K. (2006). Emotion Recognition in Spontaneous Speech Using GMMs. *Proceedings of the ICSLP Interspeech '06*, str. 809-812.

[34] Nogueiras, A., Moreno, A., Bonafonte, A., Marino, J. B. (2001). Speech Emotion Recognition Using Hidden Markov Models. *Proceedings of the Eurospeech '01*.

[35] Nwe, T. L., Foo, S. W., De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Elsevier Speech Communications Journal*, Vol. 41, No. 4, str. 603-623.

[36] Picard, R. W. (1995). Affective Computing. *MIT Media Laboratory Perceptual Computing SectionTechnical Report*, No. 321.

[37] Polzin, T. S. (2000). Verbal and non-verbal cues in the communication of emotions. *Proceedings of the ICASSP '00*, str. 2429-2432.

[38] Rong, J., Li, G., Chen, Y. P. (2009). Acoustic feature selection for automatic emotion recognition from speech. *Information processing and management*, Vol. 45, str. 315-328.

[39] Saito, S., Itakura, F. (1966). A theoretical consideration of statistically optimum methods for speech spectral density. *Report No. 3107* (in Japanese), Electrical Communication Laboratory, NTT, Tokyo.

[40] Scherer, K. R. (1979). Nonlinguistic Vocal Indicators of Emotion and Psychopathology. *In Emotions in personality and psychopathology*, Izard, C. E. (ur.), str. 495-529.

[41] Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, Vol. 99, str. 143-165.

[42] Scherer, K. R., Johnstone, T., Klasmeyer, G. (2003). *Vocal expression of emotion*, Handbook of affective sciences, str. 433-456.

[43] Schroder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. *Proceedings of the Eurospeech 2001*, Aalborg, str. 87-90.

[44] Schuller, B., Rigoll, G., Lang, M. (2003). Hidden Markov Model-Based Speech Emotion Recognition, *Proceedings of the ICASSP '03*, str. 401-404.

[45] Schuller, B., Rigoll, G., Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, *Proceedings of the ICASSP '04*, str. 577-580.
[46] Schuller, B., Villar, R. J., Rigoll, G., Lang, M. (2005). Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition, *Proceedings of the ICASSP '05*, str. 325-328.
[47] Schuller, B., Reite, S., Rigoll, G. (2006). Evolutionary Feature Generation in Speech Emotion Recognition, *Proceedings of the ICME '06*, str. 5-8.
[48] Schuller, B. et al. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge, *Speech Communication*, Vol. 53, No. 9, str. 1062-1087.
[49] Severin, F., Bozkurt, B., Dutoit, T. (2005). HNR extraction in voiced speech, oriented towards voice quality analysis, *Proceedings of the eUSiPcO*.
[50] Shen, P., Changjun, Z., Chen, X. (2011). Automatic speech emotion Recognition using support vector machine, *Proceedings of the ICEMEIT '11*.
[51] Stevens, K., Hanson, H. (1995). Classification of glottal vibration from acoustic measurments, *Vocal Fold Pshysiology*, str. 147-170.
[52] Sun, K., Yu, J., Huang, Y., Hu, X. (2009). An improved valence-arousal emotion space for video affective content representation and recognition, P*roceedings of the IEEE International Conference on Multimedia and Expo*, ICME '09, str. 566-569.
[53] Tahon, M., Degottex, G., Devillers, L. (2012). Usual voice quality features and glottal features for emotional valence detection, *Proceedings of Speech Prosody*, Shanghai, China.
[54] Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT), *Speech coding and synthesis*, Vol. 495.
[55] Tao, J., Tieniu, T. (2005). Affective Computing: A Review, *Affective Computing and Intelligent Interaction*, Springer, str. 981-995.
[56] Ververidis, D., Constantine, K. (2005). Emotional speech classification using Gaussian mixture models, *IEEE International Symposium on Circuits and Systems*.
[57] Ververidis, D., Constantine, K. (2006). Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections, *Proceedings of the XIV European Signal Processing Conference*.
[58] Watson, D., Tellegen, A. (1985). Toward a consensual structure of mood, *Psychological Bulletin*, Vol. 98, No. 2, str. 219-235.
[59] Wollmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., Cowie, R. (2008). Abandoning Emotion Classes - Towards Continuous Emotion Recognition with Modelling of Long-Range Dependencies, *Proceedings of the ICSLP Interspeech '08*, str. 597-600.
[60] Yun, S., Yoo, C. D. (2012). Loss-Scaled Large-Margin Gaussian Mixture Models for Speech Emotion Classification, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 20, No. 2, str. 585-598.
[61] (2016-05-14) http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
[62] (2016-05-14) http://www.fon.hum.uva.nl/praat/

Emotional state estimation methodology based on acoustic speech features

Branimir Dropuljić
IN2data
Data Science Company Ltd.,
Zagreb, Croatia
branimir.dropuljic@in2data.eu

Sandro Skansi
IN2data
Data Science Company Ltd.,
Zagreb, Croatia
sandro.skansi@in2data.eu

Leo Mršić
IN2data
Data Science Company Ltd.,
Zagreb, Croatia
leo.mrsic@in2data.eu

Abstract: The intelligent interaction between human and computer gained considerable interest as a subject of speech-based emotional state estimation in recent years. Estimation methodology is described in this paper by the following steps: extraction of emotional speech features, reduction of feature space and estimation of emotional states based on machine learning methods. Emotions are typically represented as discrete states (e.g. happiness, anger, fear or disgust), or as dimensions (e.g. level of valence and arousal). Classification methods are therefore used for discrete emotion recognition tasks, while regression methods are used for valence and arousal estimation. An overview of state-of-the-art acoustic speech features is presented, as well as relevant emotional speech recognition results.

Keywords: emotional state estimation, affective computing, speech signal, acoustic features, emotional speech datasets

List of figures & tables