Analiza prozodičnih značilk emocionalnega govora

Vladimir Hozjan, Zdravko Kačič, Daniel Ambruš Čeh

Fakulteta za elektrotehniko, računalništvo in informatiko
Inštitut za elektroniko
Univerza v Mariboru
Smetanova 17, SI-2000 Maribor, Slovenia
{vladimir.hozjan, kacic, daniel.ceh}@uni-mb.si

Povzetek

V članku podajamo analizo prozodičnih značilnosti emocionalnega govora. Analiza je bila izvedena na podatkovni bazi emocionalnega govora, ki je bila posneta na Fakulteti za elektrotehniko, računalništvo in informatiko v Mariboru. Emocije, ki so bile uporabljene pri snemanju baze so standardizirane v MPEG4 FAP (facial animation parameter) standardu. Baze vsebujejo govor z naslednjimi emocijami: dve nevtralni, gnus, presenečenje, veselje, strah, jeza in žalost. Bazo sta posnela profesionalna igralca (moški in ženska) v slovenskem jeziku. Izvedli smo statistično analizo visokonivojskih prozodičnih značilk (srednja vrednost poteka F0, razpon F0, srednja vrednost poteka energije, trajanje besed …). Visokonivojske prozodične značilke smo izračunali s pomočjo nizkonivojskih značilk (osnovna harmonska frekvenca (F0), energija, trajanje besed). Osnovno harmonsko frekvenco govornega signala smo izmerili s pomočjo laringografa. Primerjava rezultatov analize parametrov emocij za izolirane besede je pokazala, da obstaja v tem primeru precejšnja podobnost prozodičnih značilnosti za slovenski in angleški jezik.

Abstract

In this paper we present analysis of prosody features of emotional speech. Analysis was performed on emotional speech database that was recorded at the Faculty of Electrical Engineering and Computer Science in Maribor. Used emotions are defined in MPEG4 FAP (facial animation parameters) standard. Recorded database includes utterances with the following emotions: disgust, surprise, joy, fear, anger, sadness and two neutral. Two professional actors (male and female) recorded the database in Slovenian language. Statistical analysis was performed over high level features (average F0, F0 range, average of ΔF0, average energy, average energy of vocals, word rate, syllable rate…) that were derived from low level features (fundamental frequency (F0), energy, duration). Fundamental frequency was measured by laryngograph. Comparison of analysis results of emotional parameters for isolated words has shown that in this case similarity of prosody features for Slovenian and English language exist.

1. Uvod

Današnji sintetizatorji govora so že dosegli visoko stopnjo razumljivosti. Klub temu velikokrat še ne zvenijo popolnoma naravno. Da bi zmanjšali monotonost sitetiziranega govora je smotrno vključiti tudi vplive emocij, ki se pojavijo pri naravnem govoru. Sintetizratorji največkrat vključujejo kontrolo intonacije, ritma in jakosti. Intonacija, ritem in jakost so parametri prozodičnih značilnosti govora (Montero idr., 1998) .

Pri govorni komunikaciji med ljudmi imajo veliko vlogo tudi emocije izražene v govoru. Eksperiment prepoznavanja emocij v govoru, v katerem je sodelovalo 40 normalno slišečih ljudi, je pokazal, da je bila uspešnost prepoznavanja emocij v govoru (veselje, žalost, dva tipa jeze (vroča in hladna), in nevtralni govor) v povprečju 83%. Najbolj prepoznavna je bila žalost (96% pravilnost razpoznavanja), sledijo »vroča« jeza (86%), veselje (85%), »hladna« jeza (76%) in nevtralni govor (78%). Velikokrat so poslušalci prepoznali govor s »hladno« jezo kot nevtralno emocijo (14%), nevtralni govor pa kot emocijo žalosti (21%) (Pereira, 1996).

S ciljem, da določimo nabor značilk emocionalnega govora, ki so pomembne za razločevanje različnih emocij, smo izvedli študijo o vplivu emocij na prozodične značilke.

Ugotovitve, ki so bile podane v literaturi dosedaj, kažejo na dejstvo, da je za emocijo jeze značilno povečanje srednje vrednosti poteka osnovne harmonske frekvence (F0), razpona F0, spremenljivosti F0 in srednje vrednosti jakosti. Nekaj dokazov je tudi o tem, da se pri tej emociji poveča energija visokih frekvenc. Opažen je padajoč potek osnovne harmonske frekvence in povečana stopnja artikulacije. Stopnjo artikulacije določimo z deljenjem števila zlogov s celotnim trajanjem izgovarjave, brez upoštevanja premorov. Bolj natančno lahko določimo stopnjo artikulacije z merjenjem dolžine zvočnih segmetov (Pittam & Scherer, 1992).

Za emocijo veselja je značilno povečanje srednje vrednosti poteka F0, razpona poteka F0, spremenljivost poteka F0 in povečanje srednje vrednosti jakosti. Povečana je tudi energija visokih frekvenc.

Pri žalostnem govoru je značilno zmanjšanje srednje vrednosti poteka F0, razpona F0, spremenljivost F0, zmanjšanje energije visokih frekvenc, stopnje artikulacije in padajoč potek F0. Spremeljivost poteka osnovne harmonske frekvence je tudi zelo majhna. (Pittam & Scherer, 1992)

Scherer in Pittam sta nakazala, da je povečanje srednje vrednosti poteka F0, razpon F0, spremeljivost F0 in povečanje srednje vrednosti intenzitete in energije visokih frekvenc posledica razburjenja govorca. Veselje in jezo obravnavata kot emociji z visoko stopnjo razburjenosti govorca, žalost in nevtralni stil govora pa kot emociji z nizko stopnjo razburjenosti.

Vse zgoraj omenjene raziskave so bile izvedene za angleški jezik.

Cilj izvedene raziskave je analiza parametrov emocij v slovenskem govoru in njihova primerjava s parametri za angleški govor. Poleg tega smo želeli določiti značilke emocionalnega govora, ki različne emocije v govoru razlikujejo med seboj.

2. Slovenska baza emocionalnega govora

Slovensko bazo emocionalnega govora sestavljajo posnetki dveh profesionalnih igralcev (moški in ženska). Izbrane emocije so skladne s standardiziranimi v MPEG4 standardu in sicer znotraj FAP (facial animation parameters) protokola. To je protokol, ki opisuje parametre za animacijo obraza. FAP protokol določa parametre za šest izrazov na obrazu.

Jeza	An
Gnus	Dis
Strah	Fe
Veselje	Joy
Žalost	Sa
Presenečenje	Su
Nevtralno hitro	Nf
Nevtralno počasi	Ns

Tabela 1: Seznam uporabljenih emocij in njihove okrajšave.

Bazo tako sestavlja govor posnet v šestih emocijonalnih stanjih in dveh nevtralnih. Emocije so naštete v tabeli 1. Pri oceni emocij potrebujemo referenco. To sta v našem primeru dva nevtralna stila govora. Počasen stil govora je umirjen govor in služi kot referenca emocijam z nizko stopnjo razburjenosti (žalost in gnus). Hiter nevtralni stil govora je bolj dinamičen in smo ga uporabili kot referenco za emocije z visoko stopnjo razburjenosti govorca (jeza, veselje in strah).

Številka stavka	Vsebina korpusa
1 do 15	Izolirane besede (števke in številke)
16 do 35	Izolirane besede (fonetično bogate besede)
36 do 55	Stavki (kratki – 17 trdilnih, 3 vprašalni)
56 do 115	Stavki (srednji – 53 trdilnih, 7 vprašalnih)
116 do 135	Stavki (dolgi – 20 trdilnih, 0 vprašalnih)
136 do 190	Stavki kontekstno povezanega besedila

Tabela 2: Korpus stavkov Slovenske baze emocionalnega govora.

V tabeli 2 je podan podrobni opis korpusa stavkov baze emocionalnega govora. Korpus sestavlja 35 izoliranih besed in 155 stavkov. 100 stavkov je kontekstno nepovezanih, od tega je 20 kratkih stavkov, 60 srednje dolgih stavkov in 20 dolgih stavkov. Zadnjih 55 stavkov je kontekstno povezanih in so različnih dolžin.

Kratki stavki vsebujejo pet do osem besed, srednje dolgi stavki vsebujejo devet do trinajst besed in dolgi štirinajst do osemnajst besed.

Celotni korpus smo posneli v vseh šestih emocijah in dveh nevtralnih stilih. Poleg govornega signala smo posneli tudi signal laringografa. Ta podaja potek gibanja glasilk v grlu. Na ta način izmerimo točne vrednosti osnovne harmonske frekvence posnetega govora.

Baza je bila posneta v studiju Centra za jezikovne tehnologije na Fakulteti za elektrotehniko računalništvo in informatiko s studijskim mikrofonom AKG 3000B in laringografom Portable Laryngograph proizvajalca Laryngograph Ltd. Frekvenca vzorčenja je znašala 48KHz ob 16 bitni kvantizaciji. Snemanje smo ponovili v razmiku 14 dni z namenom ocenitev konsistentnosti posnetih emocij v govoru. V analizi, ki jo podajamo v članku smo uporabili samo posnetke prvega snemanja.

3. Izbira značilk

Za analizo smo izbrali prozodične značilke, ki smo jih izračunali iz akustičnega govornega signala in signala laringografa.

Značilke smo razdelili na nizkonivojske in visokonivojske značilke. Kot nizkonivojske značilke smo definirali intonacijo, ritem in jakost. Visokonivojske značilke smo izračunali iz nizkonivojskih značilk s pomočjo statističnih metod.

Značilka za intonacijo je potek osnovne harmonske frekvence F0. Osnovno harmonsko frekvenco F0 smo izračunali iz signala laringografa. Iz poteka F0 smo izračunali tudi drugo intonacijsko značilko in sicer odvod $\Delta F0$.

Značilko za jakost smo predstavili kot potek energije. Energijo smo izračunali z metodo kvadratične srednje vrednosti (RMS - root mean square). Za izračun poteka energije smo uporabili kvadratno okno velikosti 10ms, s korakom pomika okna 5ms. Iz poteka energije smo izračunali tudi potek odvoda energije.

Ritem določajo značilke trajanja stavkov, zlogov, besed, fonemov, premora med besedami, hitrost izgovarjave stavkov, besed, zlogov in fonemov. Izbrali smo le trajanje besed.

Visokonivojske značilke smo izpeljali iz nizkonivojskih. Določiti smo želeli čim večji nabor visokonivojskih značilk, ki bi lahko razlikovale različne emocije.

Iz značilk za intonacijo smo izračunali naslednje visokonivojske značilke: srednjo vrednost poteka F0, standardno deviacijo poteka F0, minimalno vrednost poteka F0, maksimalno vredost poteka F0, razpon poteka F0, srednjo vrednost poteka $\Delta F0$, standardno deviacijo poteka $\Delta F0$, minimalno vrednost poteka $\Delta F0$, maksimalno vrednost poteka $\Delta F0$ in razpon poteka $\Delta F0$.

Izračunane visokonivojske značilke za jakost so: srednja vrednost poteka energije, standardna deviacija poteka energije in srednja vrednost poteka energije posamezne besede.

Za ritem smo izbrali značilke: trajanje besed in trajanje zlogov.

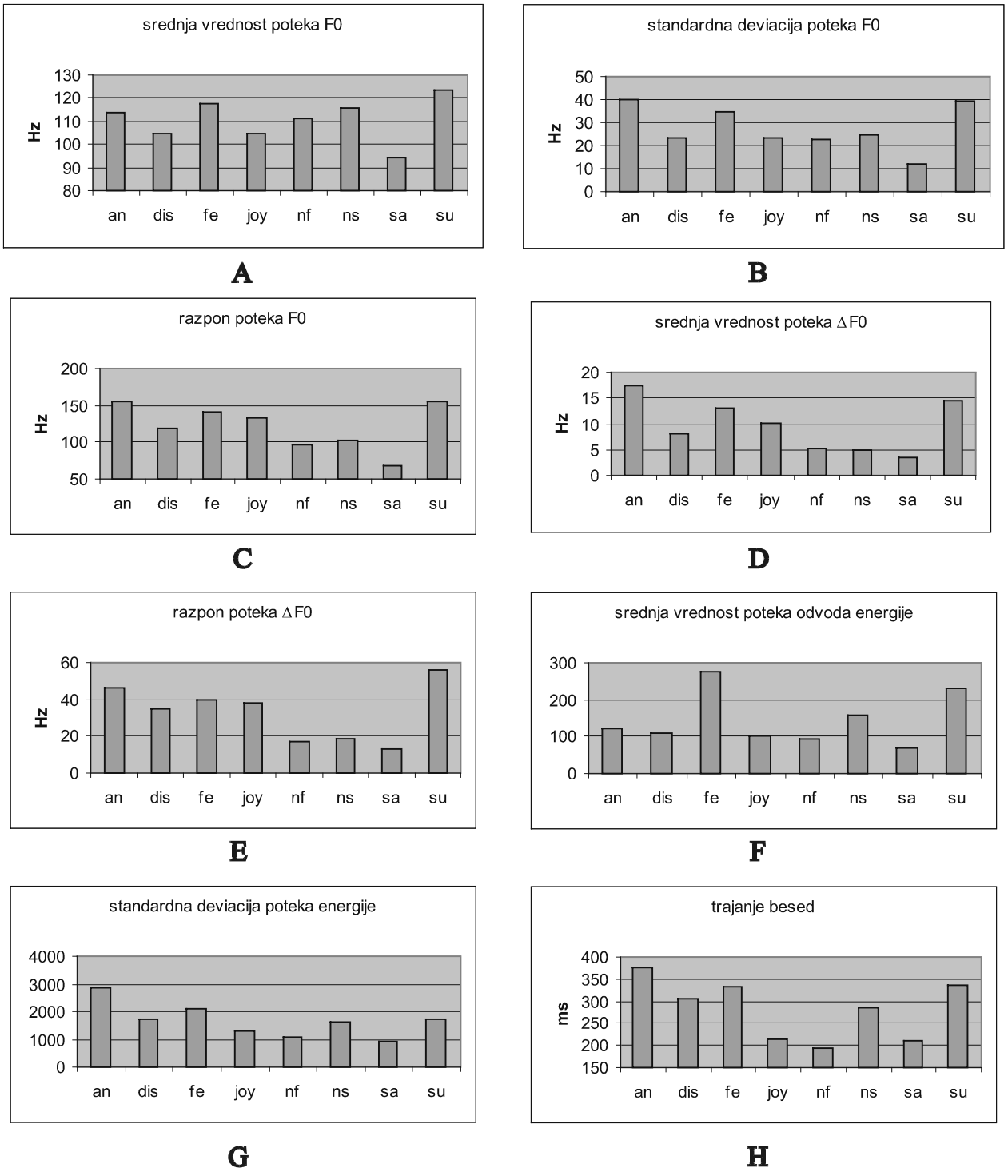
4. Analiza značilk

Analizo nizkonivojskih in visokonivojskih značilk smo izvedli za govor vseh šestih emocij in obeh nevtralnih stilov.

Pri analizi smo upoštevali prvih 35 posnetkov moškega govorca. Tako smo analizirali vpliv emocij na izgovarjavo izoliranih besed. Izolirane besede predstavljajo števke, številke in fonetično uravnotežene besede.

5. Rezultati

Dobljene rezultate visokonivojskih značilk smo predstavili v grafih in tako skušali oceniti vplive emocij na izgovarjavo. Rezultati so prikazani na sliki 1. Pri tem smo predstavili le osem visokonivojskih značilk, ki najbolj



Slika 1: Grafični prikazi visokonivojskih značilk za različne emocije. Okrajšave emocij so podane v tabeli 1.

nazorno prikazujejo značilnosti posameznih emocij in za katere lahko ugotovimo, da različne emocije v govoru najbolje razlikujejo med seboj.

6. Razprava

Primerjave angleških in slovenskih prozodičnih značilnosti govora ni pokazala velikega odstopanja med parametri posnetih emocij.

Jeza ima povečano srednjo vrednost poteka F0, standardno deviacijo poteka F0, razpon poteka F0, srednjo vrednost poteka ΔF0 in standardno deviacijo poteka energije.

Gnus se v vseh visokonivojskih značilkah ne razlikuje veliko od nevtralnega stila govora. Edina večja razlika med njima je v trajanju besed. Pri gnusu je dolžina besed bistveno večja od nevtralnega govora (slika 1.H). To velja tako za hiter kot za počasen nevtralni govor. Te emocije

ne moremo primerjati z analizami drugih avtorjev, ker ni na voljo primerljivih analiz.

Strah je v veliki meri podoben emociji jeze. Ima povečano srednjo vrednost poteka F0, standardno deviacijo poteka F0, razpon poteka F0 in srednjo vrednost poteka ΔF0. Strah se od drugih emocij razlikuje predvsem po srednji vrednosti poteka odvoda energije, ki je zelo povečan. To lastnost lahko vidimo na sliki 1.F.

Veselje ima povprečne vrednosti jakosti, podobno kot nevtralni stil govora. Ima povečano variabilnost jakosti, povečan razpon poteka F0 (slika 1.C) in povečan razpon poteka ΔF0 (slika 1.E).

Žalost je emocija, ki jo je najlažje razlikovati. Ima najmanjšo srednjo vrednost poteka F0, standardno deviacijo F0, razpon poteka F0, srednjo vrednost poteka ΔF0 in razpon poteka ΔF0.

Prozodične visokonivojske značilke za presenečenje so zelo podobne jezi in strahu. Največja razlika med strahom in jezo ter presenečenjem je v srednji vrednosti poteka F0, kjer ima presenečenje največjo vrednost.

Počasen nevtralni in hiter nevtralni govor se med seboj bistveno ne razlikujeta. Največja razlika je opazna v trajanju besed, kakor smo tudi pričakovali. Glede na druge emocije ima nevtralni govor povprečne vrednosti vseh visokonivojskih značilk.

7. Sklep

Iz opravljene analize emocionalnega govora lahko sklepamo, da ima slovenski jezik na nivoju izgovarjave izoliranih besed v določenih emocionalih stanjih podobne značilnosti kot angleški jezik. Dobljeni rezultati kažejo na podobnost značilk za jezo, strah, veselje in nevtralni govor.

Rezultati analize potrjujejo dejstvo, da človek pri prepoznavanju emocij iz govora uporablja tudi prozodične značilnosti govora. To dejstvo lahko podkrepimo z emocijo žalosti, ki jo človek najlažje razpozna. Analiza emocij je pokazala, da ima žalost najbolj različne prozodične značilnosti govora in je kot takšna najlažje razpoznana.

8. References

Pereira C., Watson C., (1998). Some Acoustic Characteristics of Emotion, In Proceeding of the Fifth International Conference on Spoken Language Processing, Sydney

Pereira C., (1996) "Angry, happy, sad or plain neutral? The identification of vocal affect by hearing-aid users", *In Proceeding of the Sixth Australian International Conference on Speech Science and Techonogy*, Adelaide

Noad J.E., Whiteside S.P., Green P.D., (1997). A macroscopic analysis of an emotional speech corpus, *In the Proceeding of the Eurospeech '97*

Cahn J.E., (1990). Generating Expression in Synthesised Speech, Magisterska naloga, *Massachusetts institute of technology*

Montero J.M., Gutierrez-Arriola J., Palzuelos S., Enriquez E., Agilera S., Pardo J.M., (1998). Emotional Speech Synthesis: From Speech Database to TTS, In the Proceedings of Fifth International Conference on Spoken Language Processing, Sydney

Pittam J., Scherer K., (1992). The encoding of affect: a review and direction for future research. *In Proceeding of the Fourth Australian International Conference on Speech Science and Technology,* Brisbane