UDK 811.163.42'342.1
159.932:534
Izvorni znanstveni rad

Juraj Bakran, Vlasta Erdeljac i Nikolaj Lazić
Filozofski fakultet, Zagreb
Hrvatska

MODELIRANJE TEMELJNIH INTONACIJSKIH OBLIKA

SAŽETAK

U Mbrola TTS (Text To Speech (tekst u "govor")) sustavu (Bakran i Lazić, 1998) osnovna frekvencija (f0) mora se zadati eksplicitno i relativno precizno. S obzirom na to da za hrvatski nema nikakvih eksplicitnih podataka o intonacijskim oblicima, metodom "analysis by synthesis" određene su veličine frekvencijskih pomaka f0 koje su potrebne da bi se odvojile granice hipotetskih osnovnih intonacijskih kategorija: "silazno", "ravno" i "uzlazno". Stimulusi, imitacija izolirano izgovorenog neutralnog samoglasnika s trajanjem od 400 ms, kreirani su Klattovim sintetizatorom.

Premda je percepcija varijacija f0 kontinuirana (za negovorne i neglazbene stimuluse), u provedenom pokusu uspješno smo izdvojili navedene tri pretpostavljene kategorije. Sasvim u skladu s teorijom o "govornom" načinu slušanja pokazalo se da se percepcijske kategorije ne poklapaju sa stvarnim oblicima varijacije frekvencije osnovnog tona, nego su zbog "prirodnog oblika" izgovorne cjeline pomaknute prema silaznom ekstremu.

Ključne riječi: *intonacijski oblici, psihoakustika, hrvatski jezik*

UVOD

Zadatak identifikacije pretpostavlja percepciju granice među kategorijama. Diskriminacija, nasuprot tomu, ima zadatak ustanoviti samo različitost ili istovjetnost dvaju signala: npr. emitiraju se tri stimulusa za redom, a slušatelj samo mora odgovoriti jesu li oni identični ili različiti. U idealnom slučaju uspješnost diskriminacije doseže maksimum upravo na mjestima koja predstavljaju granice kategorija, a pada do razine slučajnosti unutar iste kategorije. U novijim pokusima Remijsen i van Heuven (1999) pokazuju da je preciznost diskriminacije na granici među kategorijama intonacijskih oblika veća negoli unutar pojedine kategorije, čime se dokazuje kategorično načelo percepcije intonacijskih oblika.

Kad je prisutan zvuk sličan govornom zvuku, percepcija je dihotomna - takav se signal percipira ili ne percipira kao govor. Slušatelj ne mora prije početka signala biti spreman na govorni zvuk. Taj se govorni način percepcije uključuje ako u signalu ima akustičkih osobina primjerenih govoru. To uključenje govornog načina percepcije ne ovisi o volji slušatelja. Kad se sluša prirodni govor, čovjek nije u stanju negovorno slušati akustička svojstva. Isto tako, ako se sluša sintetički govor, ne sluša se niz šumova, tonova, "klikova" i stanki.

Općeprihvaćena je činjenica da intonacijski oblici imaju lingvistički distinktivnu funkciju. Upitni ili deklarativni smisao rečenice često se izražava isključivo različitim intonacijskim oblikom. Što se izvedbe tiče, moguće je zamisliti kontinuirani prijelaz od jednog intonacijskog ekstrema do drugoga. Dosadašnji pokusi pokazali su da se u jednom dijelu kontinuuma može uočiti nagli prijelaz u drugu kategoriju. Ta se pojava prikazuje izraženom "S" krivuljom. Međutim, ipak ne bi trebalo reći da je percepcija tonske visine kategorijska, nego se varijacije intonacije interpretiraju kategorijski kad se radi o tzv. "govornom (jezičnom) slušanju" (Ladd i Morton, 1997).

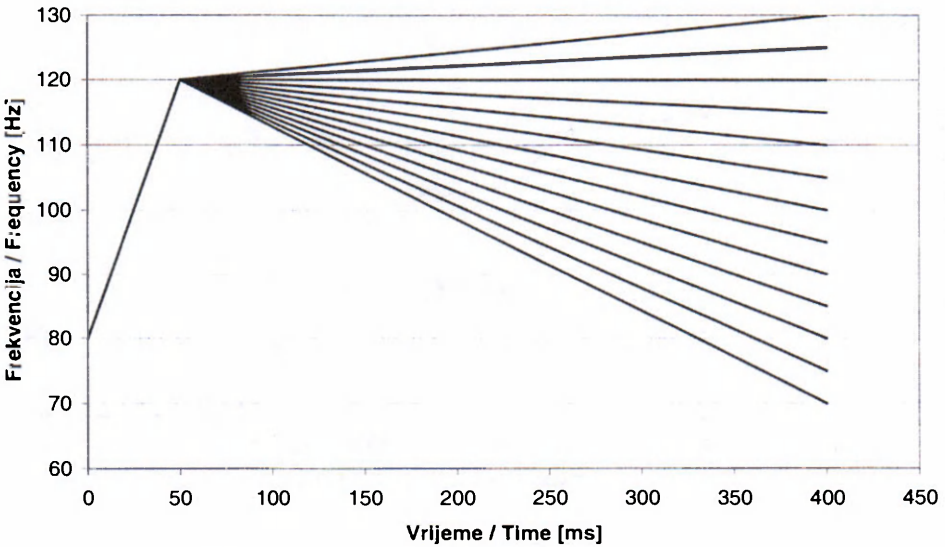
METODA

Klattovim sintetizatorom (Klatt, 1980) kreiran je neutralni vokal (*schwa*) tako da zvuči kao da je prirodno izolirano izgovoren. U 13 različitih stimulusa od po 400 ms njihova trajanja primijenjeni su različiti oblici krivulje fundamentalne frekvencije.

Za sve stimuluse f0 raste u prvih 50 ms od početnih 80 do 120 Hz, a zatim u sljedećih 350 ms f0 se linearno mijenja od 120 Hz prema završnoj točki intonacije, koja se nalazi u dijapazonu od 70 Hz do 130 Hz s korakom od 5 Hz. Shema kretanja f0 u 13 ipitivanih stimulusa prikazana je na slici 1.

Ispitanici, 74 studenta fonetike, slušali su stimuluse u četiri različita redoslijeda i procjenjivali kakva im se čini intonacija pojedinih stimulusa: da li je silazna, ravna ili uzlazna. Slušatelji su bez poteškoća prihvatili ovako formuliran zadatak, a samo rijetki su uočili da stimulusi nisu izgovoreni, nego su

sintetizirani.



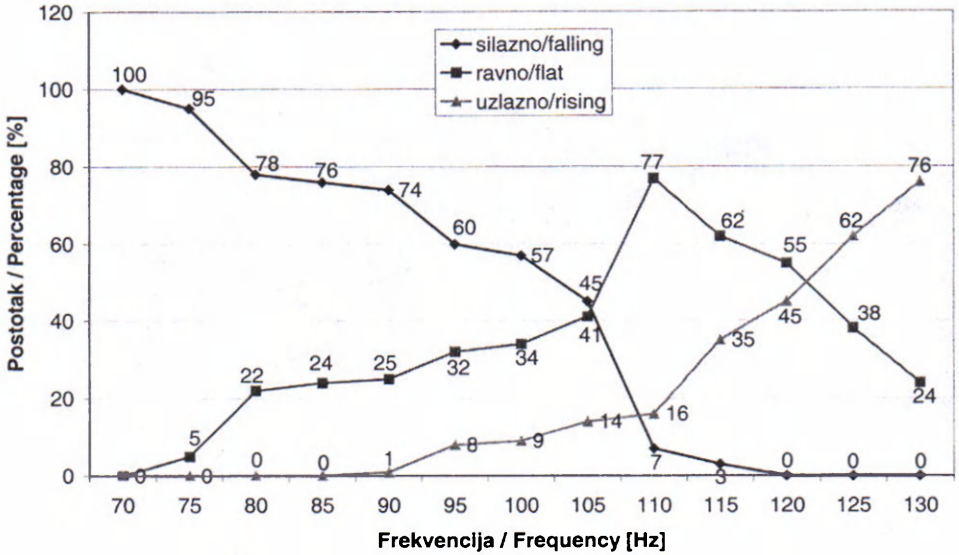
Slika1. Shema intonacijskog oblika stimulusa
Figure 1. Stimulus intonation contour

Na apscisi unešeno je vrijeme u milisekundama, a na ordinati frekvencija osnovnog tona sintetiziranog zvuka u Hz.

REZULTATI

Odgovori ispitanika prikazani su na slici 2.

Samo stimulus koji završava na 70 Hz postigao je 100%-tnu procjenu da je intonacija silazna. Već kod sljedećega pojavljuje se nekoliko odgovora da se radi o ravnoj intonaciji. Linija procjene silaznosti spušta se do nule u stimulusu kojem f0 završava na 120 Hz. Linija silaznosti siječe se s linijom koja predstavlja postotak procjene "ravno" otrprilike na 105 Hz. To se mjesto može smatrati granicom između tih dviju kategorija "silazno" i "ravno". Granica između kategorija "ravno" i "uzlazno" je na sjecištu na otprilike 123 Hz. Maksimum procjene "ravno" nalazi se na 110 Hz (prema asimetrično položenim susjednim rezultatima treba pretpostaviti da je pravi maksimum nešto više). Ekstremni stimulus ispitivanog dijapazona, onaj koji završava na 130 Hz, nije procijenjen 100% uzlaznim već pobuđuje 24% odgovora "ravno".



Slika 2. Postotak prepoznavanja silazne, ravne i uzlazne intonacije
Figure 2. Intonation identification

Na apscisi označena je završna frekvencija osnovnog tona sintetiziranih stimulusa u Hz. Na ordinati je postotak odgovora za svaku od predloženih kategorija.

RASPRAVA

Na temelju prikazanih rezultata procjene najprije možemo zaključiti da je upotrijebljen dijapazon varijacija f0 obuhvatio sve tri kategorije "silazno", "ravno" i "uzlazno", ali da nije obuhvatio granice prihvatljivih varijacija. Indikacija za ovakav komentar jest dvosmislenost stimulusa s krajem na 130 Hz, koji je za samo 76% ispitanika "uzlazan", a provocira i 24% odgovora "ravno". Niti ta kategorija "ravno" nije nedvosmisleno odvojena. Uz stimulus koji provocira najveći broj odgovora "ravno" nalazi se istovremeno 7% odgovora "silazno" i 16% odgovora "uzlazno". Zanimljivo je mjesto maksimuma procjene "ravno" koje je otprilike 10 Hz niže od onoga što bi bilo fizički ravno (120 Hz). Ovaj podatak ilustrira tipično govorni način slušanja. S obzirom na to da je diferencijalni prag za ovakvu vrstu stimulusa oko 3 Hz, taj se pomak ne može smatrati slučajnim. Procjena varijacije f0 za negovorni stimulus varirala bi oko

središnje vrijednosti 120 Hz. Prema dobivenim rezultatima može se zaključiti da za uspješno kodiranje silazne, ravne ili uzlazne intonacije treba odabrati frekvencijske pomake bliže ekstremima (prema slici 2) jer svako približavanje granicama kategorija unosi sve veći postotak dvosmislenih interpretacija. Stupanj poželjnih varijacija može se unaprijed odrediti pristajanjem na neku određenu razinu dvosmislenosti.

To se može prikazati i s odnosima završne frekvencije osnovnog tona maksimalnog raspoznavanja "ravne" intonacije (koja se nalazi 10 Hz ispod početne vrijednosti) i granica na kojima se procjenjuju kao jedna od dvije skupine (u ovom slučaju 15 Hz ispod početne vrijednosti za procjene "silazno"-"ravno" i 0 Hz u slučaju "ravno"-"uzlano").

Ta razlika vrijedi za slučaj odabrane frekvencije osnovnog tona, koja se može promatrati kao frekvencija muškoga glasa. Prenošenje intonacijskog oblika u drugo frekvencijsko područje predstavlja problem zbog različite percepcije raspona frekvencije osnovnog tona u različitim dijelovima frekvencijske ljestvice.

U akustici, frekvencija se izražava uglavnom hercima (Hz). U glazbi, različiti se frekvencijski odnosi izražavaju intervalima različite veličine (ton, poluton, terca, oktava) što je zapravo logaritamska ljestvica. U psihoakustici, najprije je predložena mel-ljestvica i kasnije bark-ljestvica (Traunmuler, 1990). Najnovija modifikacija bark-ljestvice je takozvana ERB ljestvica (*equivalet rectangular bandwidth rate*, Patterson, 1976). Bark-ljestvica je do 500 Hz linearna, a na višim frekvencijama logaritamska. ERB ljestvica ima uže kritične pojaseve od bark ljestvice (posebno na nižim frekvencijama) i nije ni linearna ni logaritamska. Postavlja se pitanje koja je od tih ljestvica primjerena prikazivanju intonativnih oblika tako da se zadovolji uvjet da ista numerička vrijednost ima isti učinak u smislu percepcije intonacije. Za modeliranje intonacije TTS sustava to je vrlo važno jer se ista informacija (model) mora moći primijeniti na različitim frekvencijama osnovnog tona. Na primjer, frekvencijska razlika od 120 do 180 Hz, ako se primijeni na početnu frekvenciju od 240 Hz, u linearnom sustavu (Hz) rezultirala bi s 300 Hz. U logaritamskom sustavu, bila bi 360 Hz, a izražena istim brojem ERB-a rezultirala bi s 325 Hz. Budući da je bark-ljestvica do 500 Hz linearna, Hermes i van Gestel (1991) dokazali su da je za kodiranje intonativnih informacija primjerena ERB ljestvica. Transformacija od frekvencijske ljestvice u ERB i obratno računa se prema:

$$E = 16{,}7 \cdot \log(1 + \frac{f}{165{,}4})$$

$$f = 165{,}4 \cdot (10^{\frac{E}{16{,}7}} - 1),$$

gdje su: E – frekvencija uzražena u ERB-ima, f– frekvencija uzražena u Hz.

To znači da se mora održati isti razmak na ERB ljestvici da bi intonacija zvučala s istim rasponom. U našem primjeru se raspon od 120 do 105 Hz (razlika 15 Hz ili 0,3916 na ERB ljestvici) prenosi na raspon 240 do 219 Hz (razlika 21 Hz ili 0,3916 na ERB ljestvici). Na ovom frekvencijskom području bi frekvencija od 219 Hz bila bi granica između "silazno"-"ravno".

Za sintetiziranje prozodije govornih signala što bližih prirodnom govoru potrebno je "ravnu" intonaciju napraviti blago silaznom i paziti na transpozicije intonacijskog oblika zbog različite percepcije frekvencijskih raspona u slušanju.

REFERENCIJE

Bakran, J. i Lazić, N. (1998). Fonetski problemi difonske sinteze hrvatskoga govora. *Govor* **?**, 2, 103-116.

Hermes, D. J. and van Gestel, J. C. (1991). The frequency scale of speech intonation. *J. Acoust. Soc. Am.* **90**, 97-102.

Klatt, D. H. (1980) Software for cascade/parallel synthesizer. *J. Acoust. Soc. Am.* **67**, 971-975.

Ladd, D. R. and Morton, R. (1997). The perception of intonational emphasis: continuous or categorical? *Journal of Phonetics* **25**, 313-342.

Patterson, R. D. (1976). Auditory filter shapes derived with noise stimuli. *J. Acoust. Soc. Am.* **59**, 640-654.

Remijsen, B. and van Heuven,V. J. (1999) Gradient and Categorical Pitch Dimensions in Dutch: Diagnostic Test, *Proceedings ICPhS99, San Francisco, 1865-1868.*

Traunmuller, H. (1990). Analytical expression for the tonotopic sensory scale, *J. Acoust. Soc. Am.* **88**, 97-100.

Juraj Bakran, Vlasta Erdeljac and Nikolaj Lazić
Faculty of Philosophy, Zagreb
Croatia

MODELLING BASIC INTONATION FORMS

SUMMARY

In Mbrola TTS (Text To Speech) system, the f0 must be set explicitly and with relative precision. Given the fact that for Croatian there are no explicit data on intonation forms, the "analysis by synthesis" method was used to determine what sizes of frequency f0 shifts were required to separate the limits of hypothetically basic intonation categories: falling, flat, rising. The stimuli, imitation of separately uttered neutral vowels in duration of 400 ms were all created by means of Klatt's synthesizer. Although perception of f0 variations is a continued process (for non-speech and non-musical stimuli), in our experiment we successfully isolated the three formerly set categories. In complete accordance with the theory of "speaking-like" listening, it appeared that perceptive categories did not correspond to the actual forms of frequency variations of the basic tone, but were, due to the "natural form" of the pronunciation unit, shifted towards the falling extreme.

Key words: *intonation forms, psychoacoustics, the Croatian Language*