UDK 808.62-41:534
Izvorni znanstveni rad

Prihvaćeno 01.12.1998.

Juraj Bakran i Nikolaj Lazić
Filozofski fakultet, Zagreb
Hrvatska

FONETSKI PROBLEMI DIFONSKE SINTEZE HRVATSKOGA GOVORA

SAŽETAK

Termin "sinteza govora" obuhvaća niz postupaka kojima se s ograničenim, što manjim utroškom memorije, ozvučava, prevodi u kvazi govoreni oblik, po mogućnosti neograničen korpus pisanog jezika. Jedna je od metoda sinteze nizanje unaprijed snimljenih, u digitalnom obliku pohranjenih, govornih elemenata. Kod toga, temeljni je problem odabir govornih elemenata (rečenice, riječi, slogovi, fonemi) koji će se zadanim redoslijedom kombinirati u nizove. Problem koartikulacije i interpolacije u velikoj se mjeri rješava odabirom difona. Difonom se naziva govorni segment od sredine jednog do sredine slijedećeg alofona (Fujimura et al. 1977.) Na taj su način sačuvani prirodni tranzijenti, izbjegava se diskontinuitet na spoju i interpolacija postaje nepotrebna.

U fonetskom laboratoriju Faculte Politechnique de Mons, grupa autora (Dutoit et al., 1996a, 1996b) razvila je program za difonsku sintezu koji za realizaciju koristi izvorne snimke repertoara difona nekog konkretnog jezika. Program je razvijen u okviru projekta pod nazivom MBROLA i preko Interneta je dostupan potencijalnim korisnicima i autorima baze difona. Krajnji cilj projekta je poticanje akademskog istraživanja u području sinteze govora u godinama koje dolaze, posebno na planu istraživanja prozodije sintetiziranog govora.

U radu se opisuje postupak kreiranja baze difona za hrvatski standardni govor. S obzirom na to da baza difona mora sadržavati sve moguće prijelaze jer u kontinuiranom govoru na granici riječi svi fonemi mogu doći u neposredni kontakt, trebalo je sastaviti takav tekstovni korpus u kojem svih 30 hrvatskih fonema dolaze u kontakt, svaki sa svakim. To je ujedno i prilika da se realiziraju svi mogući alofoni, oni koji se u opisu nekog govora opisuju, kao i oni koji izmiču pažnji, kako fonetičara tako i govornika. Uz pretpostavku da su sve glasovne promjene (ozvučavanje, obezvučavanje, djelomične asimilacije...) zapravo uzrokovane fonetskim kontekstom, uputa govorniku da ne artikulira suviše pažljivo osigurala je pojavljivanje alofona.

Ključne riječi: *difoni, govorna sinteza, hrvatski jezik*

UVOD

Govorna je sinteza na suprotnom polu od automatskog prepoznavanja govora jedan od osnovnih interesa tehnologije govora. Termin obuhvaća niz postupaka kojima se s ograničenim, što manjim, utroškom memorije ozvučava, prevodi u kvazi govoreni oblik, po mogućnosti neograničen korpus pisanog jezika. Jedna od metoda sinteze jest nizanje unaprijed snimljenih i u digitalnom obliku pohranjenih govornih elemenata. Pritom, temeljni je problem odabir govornih elemenata (rečenice, riječi, slogovi, fonemi) koji će se zadanim redoslijedom kombinirati u nizove. Ako se odabere rije~čkao osnovna jedinica sinteze, što se čini razumnim izborom, i ako se snimi u izoliranom obliku, postaje veliki problem nizanje takvih konkretnih riječi u veće cjeline. Osim velikog broja riječi (a u hrvatskom se taj broj zbog fleksija višestruko povećava) koje treba snimiti, problem koartikulacije i rečenične prozodije teško je riješiv. Moglo bi ga se riješiti odabirom rečenice kao osnovnog elementa, ali se takva metoda može primijeniti samo u slučaju kad je potreban rečenični korpus relativno ograničen (u vezi s nizanjem riječi i rečenica vidi: Allen et al. 1979). Jedna od mogućih temeljnih jedinica na razini niže od riječi jesu morfemi. Njih ima u jeziku znatno manje nego riječi. U tom se slučaju postavlja kao problem segmentacija u morfeme. Osim toga potrebno je na neki način nizove morfema, njihove kontakte, izgladiti. Taj se postupak naziva interpolacijom. Odaberu li se slogovi kao temeljne jedinice (Yannakoudakis i Hutton 1987.) ili još manje fonemi, odnosno alofoni, interpolacija postaje sve važnija, jer vokalski se trakt i pripadajući akustički rezultat ne mogu naglo mijenjati. Problem koartikulacije i interpolacije u velikoj se mjeri rješava odabirom difona. Difonom se naziva govorni segment od sredine jednog do sredine sljedećeg alofona (Fujimura et al. 1977) Na taj su način sačuvani prirodni tranzijenti, izbjegava se diskontinuitet na spoju i interpolacija postaje nepotrebna.

U fonetskom laboratoriju Faculte Politechnique de Mons, Belgija, skupina autora (Dutoit et al., 1996a, 1996b) razvila je program za difonsku sintezu koji za realizaciju koristi izvorne snimke repertoara difona nekog konkretnog jezika. Program je razvijen u sklopu projekta pod nazivom MBROLA i preko Interneta je dostupan potencijalnim korisnicima i autorima baze difona. Potencijalni korisnici programa moraju pristati na uvjete autora koji ograničavaju njegovu primjenu. Program se ne može prodavati niti ugraditi u neku drugu cjelinu koja se prodaje bez odobrenja autora. Bez naplate, program se može kopirati i distribuirati. Dakle, program se ne smije koristiti u komercijalnom smislu niti za vojne svrhe. Krajnji cilj projekta je poticanje akademskog istraživanja u području sinteze govora u godinama koje dolaze, posebno na planu istraživanja prozodije sintetiziranog govora. Zainteresirani za program najbolje je da pogledaju Internet stranicu MBROLA projekta na adresi:

http://tcts.fpms.ac.be/synthesis/mbrola.html

Uz adekvatne programske dodatke i pristanak autora, pomoću ovdje opisanog sintetizatora može se zamisliti niz primijena za hendikepirane osobe kao što je sat koji govori ili čitanje napisanih (e-mail) poruka. U tom slučaju treba primijeniti komercijalnu verziju MBROLA sintetizatora koja se od Internet verzije razlikuje samo po tome što je baza podataka memorirana u kompaktnom obliku koji se dekodira u realnom vremenu.

KREIRANJE BAZE DIFONA

Jedan od najvažnijih aspekata MBROLA projekta mogućnost je jednostavnog dodavanja novih jezika i novih glasova (govornika). Tako je do sada na raspolaganju potencijalnim korisnicima već 10 jezika: engleski (britanski i američki), francuski, grčki, nizozemski, njemački, portugalski, rumunjski, španjolski, švedski... i s našim prilogom odnedavno hrvatski, od kojih neki sadrže difonske baze različitih muških i ženskih govornika, a broj novih jezika, odnosno, baza difona neprestano se povećava.

Snimanje difonske baze podataka ni u kojem slučaju nije trivijalan zadatak. Ako taj dio posla nije pažljivo napravljen, rezultat može razočarati. Formiranje baze difona ima nekoliko stadija:
- formiranje tekstualnog korpusa
- snimanje tekstualnog korpusa
- segmentiranje difona
- intenzitetsko izjednačavanje difona

Difoni su govorni elementi koji počinju u sredini jednog glasa (fonema) i završavaju na sredini sljedećega. Takav odabir govornih segmenata isključuje probleme sintetiziranja prijelaza (tranzijenata), te djelovanje koartikulacije i olakšava nizanje a istodobno štedi potrebnu memoriju u odnosu na sintezu s poluslogovima ili trifonima.

Prema tome, prvo treba ustanoviti broj glasova u jeziku. Glasovi su govorna ostvarenja fonema, a fonemi su lingvistički definirani na funkcionalnoj razini. Međutim, repertoar različitih glasova koji uključuje sve alofone nastaje zbog koartikulacijskih utjecaja u međusobnom kontaktu fonema. S obzirom na to da baza difona mora sadržati sve moguće prijelaze jer u kontinuiranom govoru na granici riječi svi fonemi mogu doći u neposredni kontakt, trebalo je sastaviti takav tekstovni korpus u kojem svih 30 hrvatskih fonema dolaze u kontakt, svaki sa svakim. To je ujedno i prilika da se realiziraju svi mogući alofoni, oni koji se u opisu nekoga govora opisuju, kao i oni koji izmiču pozornosti fonetičara i govornika. Problem uključivanja alofona riješili smo primjerenom uputom govorniku. Uz pretpostavku da su sve glasovne promjene (ozvučavanje, obezvučavanje, djelomične asimilacije...) zapravo uzrokovane fonetskim kontekstom, uputa govorniku da ne artikulira suviše pažljivo osigurala je pojavljivanje alofona. Pritom, svjesni smo da stupanj organiziranosti ili opuštenosti artikuliacije nije moguće definirati i zadati, pa to ostaje slučajan odabir govornika za vrijeme snimanja.

Da bismo sve difone smjestili u sličan kontekst, oni su umetnuti u rečenicu: "Reci .... opet". Rečenični naglasak bio je na posljednjoj riječi (opet). Logatomi koji su u rečenicu umetnuti sastavljeni su od četiri fonema tako da prijelaz između srednja dva predstavlja difon koji nas zanima, a prvi i zadnji dodani su tako da po kriteriju vokal-konsonant budu različiti od onih s kojima su u dodiru. Pritom, konsonant je uvijek bio /t/, a vokal je uvijek bio /a/. Tako na primjer, logatom /adot/ sadrži spoj fonema /do/, a logatom /azga/ sadrži spoj fonema /zg/. Na taj način minimalna, ali potpuna lista, ako se isključi kontakt dvaju istih fonema, sadrži 870 logatoma. Tomu treba dodati 30 početnih i 30 završnih glasova.

SNIMANJE KORPUSA

Odabrani korpus čitao je profesionalni govornik (zagrebački glumac Rene Medvešek) u studijskim uvjetima. Snimka je, prema uputama autora programa, pohranjena u digitalnom obliku sa 16 bita i frekvencijom uzimanja uzoraka 16 kHz. Govornik je trebao čitati maksimalno monotono. To je postignuto tako da je za vrijeme vježbe govornik ugodio sebi najprirodniji registar i čim bi odstupio od jednom odabrane tonske visine, snimanje je zaustavljeno i ponovno je snimkom potaknut na imitiranje početne monotone intonacije. Zbog ovih neprirodnih zahtjeva, govornik mora biti iskusan profesionalac. Ambijentalna buka i reverberacija svedene su na minimum snimanjem u audio-studiju.

SEGMENTIRANJE KORPUSA

Snimljeni korpus okvirnih rečenica mora se segmentirati, tako da se iz svake rečenice izdvoji difon kojemu se sa svake strane ostavlja još 800 uzoraka. Takav izdvojen difon snima se u datoteku s posebnim imenom, a u jednu drugu, tekstualnu datoteku upisuju se podaci koji pobliže opisuju izdvojeni difon: njegovo ime, ukupno trajanje i trajanje do granice među fonemima izraženo u broju uzoraka. Ovo posljednje poseban je problem kod razmjerno kontinuiranih prijelaza (npr. dva vokala). Određivanje granice između glasova nužno je zato da se u postupku sinteze može produživati ili skraćivati pojedini glas, a da to ne djeluje na susjedni.

INTENZITETSKO IZJEDNAČAVANJE DIFONA

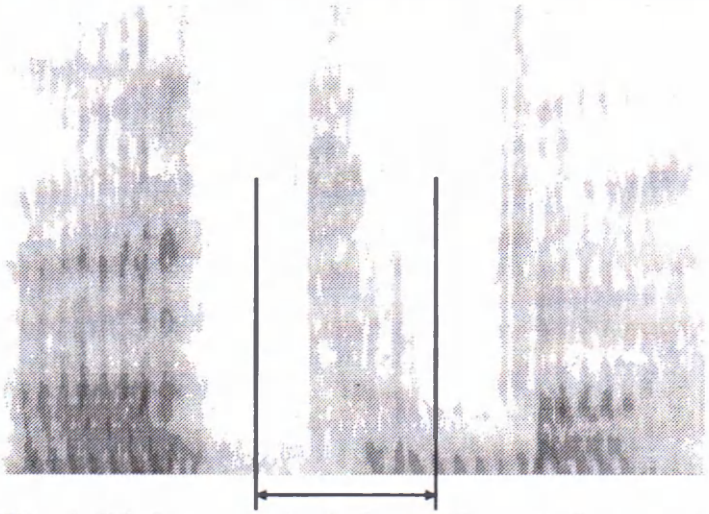
S obzirom na to da su difoni koji se nastavljaju jedan na drugi izgovoreni u različitim logatomima, oni nisu izgovoreni istom govornom snagom. Izjednačavanje difona po amplitudi je prema preporuci autora algoritma, fakultativno. Po našem mišljenju, izjednačavanje se može provesti jedino u fazi snimanja, pazeći na vršnu vrijednost naglašene (uvijek iste) riječi. Samo se na taj način može sačuvati inherentni intenzitet pojedinih glasova (difona). Ipak, valja

spomenuti da se u fazi prilagodbe segmentiranih difona MBROLA formatu obavlja i jedna vrsta amplitudnog izjednačavanja. To je potrebno jer se kod nizanja difona susreću dva dijela istovrsnoga glasa i amplitudni diskontinuitet ne smije biti čujan.

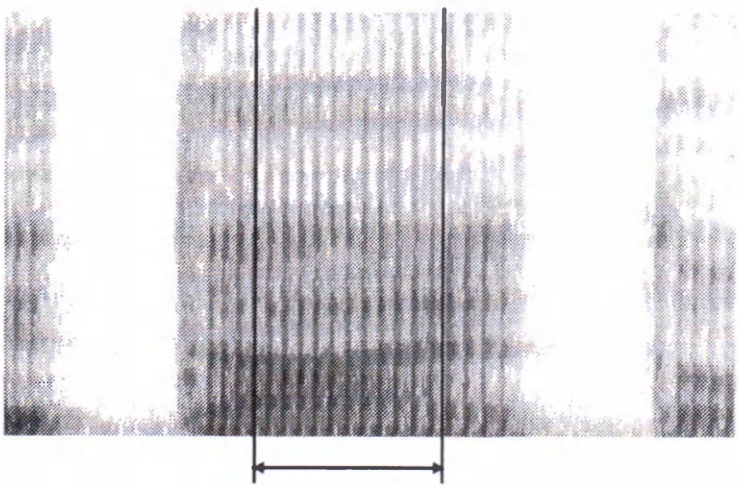
PROBLEMI SEGMENTIRANJA

Unutar okvirne rečenice, već na dijagramu valnog oblika, lako se uočava segment koji nas zanima zbog kontrastnog tipa okoline. Povećavanjem mjerila prikaza (zoom) na monitoru se postavljaju samo srednja četiri fonetska segmenta. Tada treba odrediti granice segmenata. Relativno je jednostavno izdvojiti dva glasa (alofona) koji sadrže difon zbog kojega je logatom konstruiran, zato što konsonantu prethodi vokal i obrnuto, prije vokala je konsonant. Teškoća je samo odrediti granicu između /a/ i /j/. Međutim, da bi se mogla odrediti sredina prvog i drugoga glasa (alofona), "lijeve i desne" strane difona, treba odrediti i granicu između njih. U većini slučajeva kada su glasovi fonetski relativno različiti, to nije problem.

Kada valni oblik nije dovoljno jasan, za segmentiranje se koristi spektrografski prikaz. U slučaju kada su u dodiru dva vokala /ao, ou .../, ili dva slična frikativa /sS, zZ .../, akustički je oblik relativno kontinuirani prijelaz, pa se granica određuje na sredini ukupnog trajanja. Prema tome, početkom difona smatra se 25 % ukupnog trajanja, sredina je 50 %, a kraj difona je 75 % trajanja unutar okvirnog konteksta. Na spektrogramima su prikazana dva primjera segmentacije: na sl. 1. prikazan je slučaj s fonetski jasno različitim fonetskim segmentima s jasnom granicom među njima i drugi primjer, sl. 2. spoj dvaju vokala s kontinuiranim prijelazom. Strelicama su označeni dijelovi trajanja koji su izdvojeni u poseban difon. trajanja koji su izdvojeni u poseban difon.

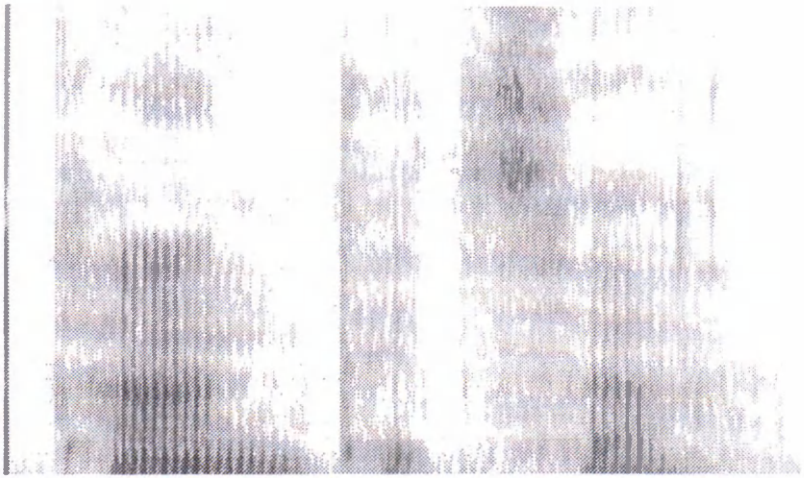


Slika 1. Primjer segmentiranja fonetski kontrastnih segmenata /atba/
Figure 1. An example of segmenting phonetically contrastive segments/atba/

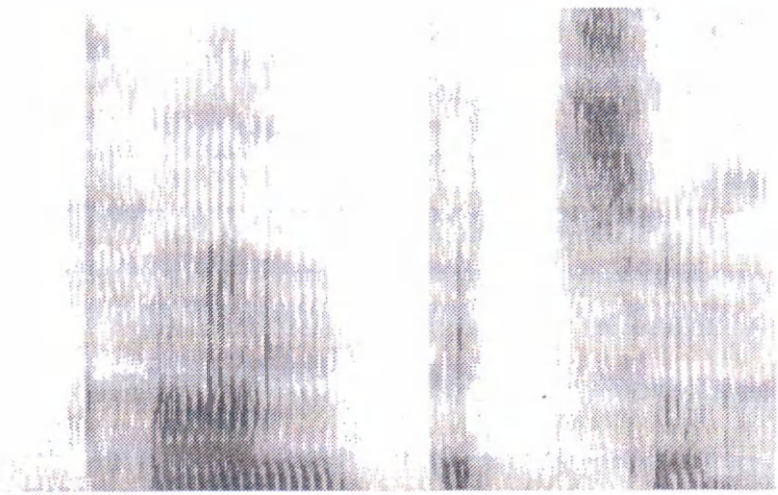


Slika 2. Primjer segmentiranja spoja dvaju vokala /taeta/
Figure 2. An example of segmenting the conjunction between two vowels /taeta/

sintetizirane verzije (sl. 4) može se primijetiti da su vokali sastavljeni iz raznih dijelova, spajanjem različitih difona.



Slika 3. Originalni izgovor riječi /kantica/
Figure 3. The original pronunciation of the word /kantitsa/



Slika 4. Rezultat sinteze fonemskog slijeda /kantica/
Figure 4. The results of the synthesis of the phonemic sequence /kantitsa/

SINTETIZIRANJE NAGLASKA

Budući da se intenzitet pojedinih zvučnih segmenata zadavanjem parametara .pho datoteci ne može regulirati, prozodija se može interpretirati samo trajanjem i tonom segmenata. To se odnosi i na razinu riječi i na razinu rečenice. Unatoč tome što su u prirodnom govoru varijacije intenziteta neprestano prisutne, pokazalo se da se zadovoljavajući stupanj razumljivosti i "prirodnosti" može postići i bez tog parametra. Dakako, nije nepoznato da u prirodnom govoru sve ove tri zvučne dimenzije: trajanje, intenzitet i visina tona, zajedno i međuovisno, u skladu s takozvanim "trade effect-om", kreiraju elemente prozodije. To znači da se isti učinak može postići različitom zastupljenošću pojedine od ovih temeljnih dimenzija zvuka. Prema tome, djelovanje povećanog intenziteta u prirodnom izgovoru može se zamijeniti produljenjem i većom tonskom visinom. To ne znači da su sve ove tri dimenzije zvuka međusobno posve zamjenjive. Ipak, čini se da je utjecaj intenziteta najlakše kompenzirati.

ZAKLJUČAK

Možda u uvodnom dijelu članka nije dovoljno istaknuto da ovaj rad ne predstavlja cjelovito rješenje sinteze hrvatskog govora već samo jednu fazu. Ono bitno što nedostaje jest način zadavanja prozodije u širem smislu riječi. U ovom dijelu osnovana je baza podataka i osigurana programska podrška za reprodukciju. Za bilo koju vrstu primjene nužno je eksplicitno zadavati trajanja segmenata i tijek frekvencije osnovnog tona. Zato u ovoj fazi rada uspješnost sinteze bitno ovisi o spretnosti onoga koji upisuje potrebne parametre. Ako se pomoću akustičke analize prirodnoga izgovora odrede parametri i unesu u .pho datoteku kojom se zadaje tekst sinteze rezultat je posve razumljiv govor i samo povremeno čuju se artefakti sinteze koji nastaju zbog nedovoljno uspješnog izjednačavanja difona u fazi pripreme i u fazi spajanja. Međutim, kako su prvi pokusi pokazali ovaj se sustav vrlo uspješno može koristiti u laboratorijskom kreiranju govornih stimulusa za raznovrsna prozodijska istraživanja jer omogućuje kontrolirano variranje trajanja segmenata i frekvencije osnovnog tona.

REFERENCIJE

Allen, J., Carlson, B., Granstrom, B., Hunnicutt, S., Klatt, D. and Pisoni, D. (1979). *Conversion of Unrestricted English Text to Speech*. Cambridge: MIT.

Bakran, J. (1996). *Zvučna slika hrvatskoga govora*. Zagreb: Ibis Grafika.

Bakran, J. i Horga, D. (1996). SAMPA za hrvatski. *GOVOR* XIII, 1-2, 99-106.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F. and Van Der Vrecken O. (1996). The MBROLA Project: Towards a Set of High-Quality Speech

Synthesizers Free of Use for Non-Commercial Purposes. *Proceedings of ICSLP'96*, Philadelphia, vol. 3, 1393-1396.

Dutoit, T. (1996). *An Introduction to Text-To-Speech Synthesis*. Kluwer Academic Publishers.

Fujimura, O. Macchi, M. J. and Lovins, J. B. (1977). Demisyllables and Affixes for Speech Synthesis. *Proceedings of the 9th Int. Congr. on Acoust.*, 51-53.

Yannakoudakis, E. J. and Hutton, P. J. (1987). *Speech Synthesis and Recognition Systems*. Ellis Horwood Limited.