Mapiranje fonema i vizema kod virtuelnog govornika na srpskom jeziku

Ana Gavrovska, *IEEE Student Member*

***Sadržaj* — Prikupljanje podataka pomoću Facial Motion Capture-a je najbolji način generisanja lica u virtuelnom prostoru putem beleženja pokreta iz realnosti. Kako je fokus istraživanja već odavno prebačen na algoritme i tehnike, a ne na samu opremu za Motion Capture, programiranje softvera seli se u manje razvijene zemlje.**

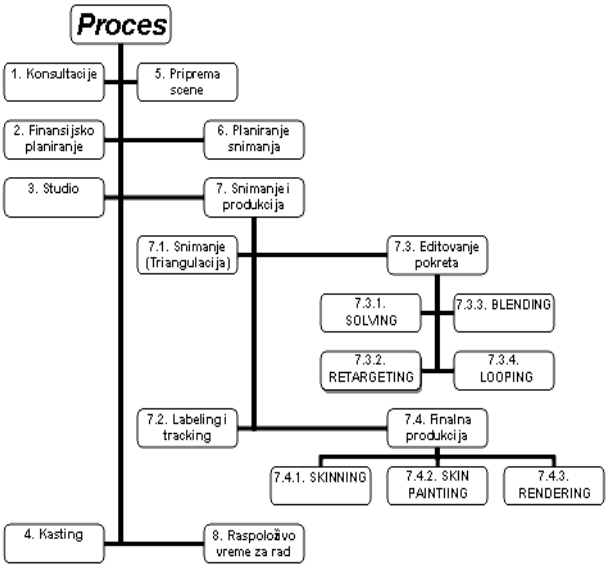
Ovde su prikazane osnove primene MPEG-4 u animaciji lica. Ukratko je objašnjena parametrizacija i specifikacija lica definisana standardom MPEG-4 FA. Od parametara animacije lica (FAPs) visokog nivoa razmatrani su vizemi i mogućnost mapiranja fonema i vizema srpskog jezika u sistemu generisanja virtuelnog govornika.

***Ključne reči* — MC, MPEG-4 FA, FAPs, FPs, vizemi, fonemi, mapiranje.**

I. UVOD

Komunikacija sa ljudskim likom je dopadljivija od standardne komunikacije sa mašinom. Ovakav vid komunikacije nosi određenu meru prisnosti. U računarskoj grafici animacija ljudskog lica je veliki izazov.

Najrealističniji pokreti lica se dobijaju beleženjem pokreta iz realnosti pomoću Motion Capture-a (MC) i prenošenjem u virtuelni prostor [1]. Najveći nedostatak ovih sistema je njihova cena. Na sl.1. su prikazani osnovni koraci u procesu MC-a [1], [2].



Sl.1. Proces Motion Capure-a.

Ana Gavrovska, Elektrotehnički fakultet u Beogradu, Srbija (telefon: 381-63-318728; e-mail: anaga@eunet.yu).

Najvažniji korak je simanje i produkcija. Na licu se postave odgovarajući markeri, koji se obeleže i prate pri kretanju samog lica. U okviru editovanja pokreta, prvo se kreira lobanja, odnosno indeksirana mreža poligona (*solving*), zatim se vrši pažljivo prenošenje snimljenih pokreta na model lica, koji može biti i karikaturalnog izgleda (*retargeting*). Više pokreta se spajaju u jedan (*blending*), a posebna pažnja se posvećuje onim pokretima koji se mogu ponavljati više puta (*looping*). U okviru finalne produkcije ,,koža se navlači'' na model lobanje u odgovarajućem softveru (*skinning*), dok se karakteristike kao što je boranje kože i određene deformacije izazvane pokretima lica doteruju naknadno (*skin painting*). *Rendering* predstavlja konačno procesiranje dobijenog 3D modela.

Facial Motion Capture predstavlja najizazovniji tip MC-a, koji će još dugo biti u žiži interesovanja. Koliko će se detaljno izraditi model zavisi od: namene samog modela, raspoložive procesorske moći, raspoloživog vremena za izradu i raznih drugih ograničenja.

MPEG-4 FA standard je ukratko objašnjen u drugoj glavi, dok su u trećoj glavi date neophodne faze u procesu generisanja virtuelnog lika.

U četvrtoj glavi je prvi put izvršeno mapiranje fonema i vizema srpskog jezika u cilju iskorišćavanja gotovih sistema u engleskom jeziku. Takođe su prikazane blok šeme koje pokrivaju moguće slučajeve.

II. RAZVOJ I OSNOVE MPEG-4 FA STANDARDA

Standardizacija lica je uvedena preko standarda MPEG-4 FA (*Facial Animation*). U okviru MPEG grupe za FBA (*Facial and Body Animation*), koja se bavila definisanjem parametara za lice i telo, grupa SNHC (*Synthetic Natural Hybrid Coding*) je radila na integraciji objekata i audia.

Sa MPEG-4 FA, osim standarda za kodiranje videa nižeg protoka bita od prethodnih, uveden je potpuno nov koncept opšteg standarda za višemedijski sadržaj sa standardizovanim formatom i značenjem podataka, dok ni enkoder, ni dekoder nisu u potpunosti definisani [3].

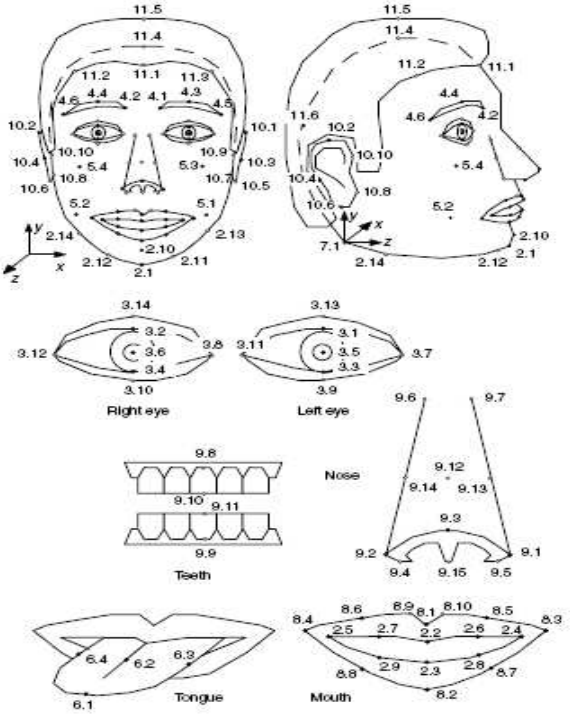
Uključen je koncept kodiranja slike zasnovan na modelu, dok je za osnovnu jedinicu kodiranja izabran AVO (audio-vizuelni objekat) upravo zbog nižeg protoka bita. Svi objekti se nezavisno koduju i stvaraju zasebne stream-ove, a za prenos se multipleksiraju zajedno sa opisom scene. Za opis MPEG-4 scene upotrebljava se BIFS (*Binary Format for Scene Description*) sa definisanim prostornim i vremenskim rasporedom AVO-a. Sami objekti opisani su pomoću VRML (*Virtual Reality*

Motion Language) jezika, korišćenjem čvorova sa definisanom rotacijom, skaliranjem i translacijom, kontruišući indeksiranu mrežu poligona.

FA ima dva dela i to FA niskog nivoa (*low-level*), koji koristi promene geometrije lica u vremenu i visokog nivoa (*high-level*), koji koristi parametre niskog nivoa. Pristup animaciji može biti različit. Najčešće se koristi metoda interpolacije (*key expression interpolation*), kojom se može omogućiti veliki broj promena izraza između dva ključna. MPEG-4 parametrizacija nije idealna. To znači da nije moguća specifikacija bilo kakvog lica sa proizvoljnim izrazom i postepenim prelazima iz jednog u drugi i to jednostavnom promenom parametara. Neophodan je kompromis između raspona mogućih izraza i jednostavnosti upotrebe baze raspoloživih ključnih izraza lica.

U MPEG-4 FA definisani su parametri za promenu geometrije lica FDPs (*Facial Definition Parameters*), parametri za definisanje osnovnih izraza lica FAPs (*Facial Animation Parameters*) i karakteristične tačke FPs (*Feature Points*) za manipulaciju nad promenama izraza. Za sve FAP-ove definisane su deformacije osnovnih izraza lica u FAT tabelama (*Facial Animation Tables*). Ovakvim tabelama ostvaruje se predvidljivost i prenosivost pokreta na različite modele.

Dovoljno malim FAPU (*Face Animation Parameter Unit*) jednicama (*iris diameter* IRISD0*, eye separation* ES0*, eye-nose separation* ENS0*, mouth-nose separation* MNS0*, mouth width* MW0*, angle unit* AU), koje su geometrijski intuitivne i definisane na neutralnom modelu lica (*Face Model in Neutral State*) postiže se nijansiranost u pomerajima lica. Na sl.2. prikazano je neutralno lice sa rasporedom 84 FP-a, koji omogućavaju manipulaciju nad FAP-ovima.



Sl.2. Raspored FP-ova [3].

FAP-ovi su izabrani kao međusobno nezavisni, ali se takođe potpuna nezavisnost ne može postići. U 10 grupa je raspoređeno 68 FAP-ova, koji mogu biti niskog i visokog nivoa apstrakcije. Ako su istovremeno definisani parametri oba nivoa, onda se u obzir uzimaju samo parametri niskog nivoa. Za FAP-ove visokog nivoa, protok od 0.3 kbps je obezbeđen, što je dovoljno za veliki broj aplikacija.

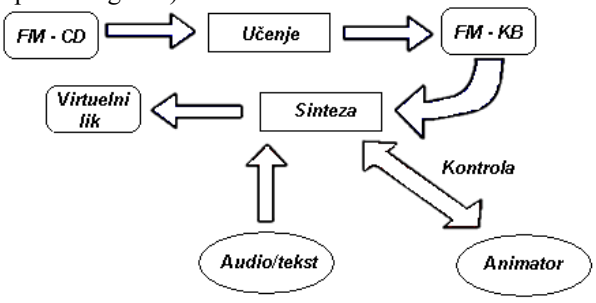
TABELA 1: GRUPE FP-OVA [3].

Grupa FP-ova	***Opis***	***Broj FAP-ova u grupi***
1.	Vizemi i izrazi	2
2.	Čeljust, brada, unutrašnja kontura usana	16
3.	Očna jabučica, zenica, očni kapci	12
4.	Obrve	8
5.	Obrazi	4
6.	Jezik	5
7.	Rotacija glave	3
8.	Spoljašnja kontura usana	10
9.	Nos	4
10.	Uši	4

FAP-ovi visokog nivoa apstrakcije su vizemi i izrazi. Vizemi su vizuelni prikaz fonema. U okviru MPEG-4 FA definisano je 14 statičkih vizema. Sa druge strane, definisano je 6 osnovnih izraza (radost, tuga, ljutnja, strah, zgražavanje, čuđenje), koji su određeni vrednostima pobude. Izrazi upotpunjuju vizeme u smislu osećajnosti, ali oni ovde neće biti razmatrani.

III. PROCES GENERISANJA VIRTUELNOG LIKA

U sistemu za generisanje virtuelnog lika neophodne su dve faze. To su faza učenja i faza sinteze. Ako na raspolaganju imamo gotovu bazu vizema FM – CD (*Facial Motion Capture Data*), nakon „učenja'' osnovnih izraza postaje FM – KB (*Facial Motion Knowledge Base*). Na raspolaganju korisnik ima bazu ključnih vizema. Ulaz u sintezu virtuelnog lika je FM-KB i tekst (za *text-to-speech* aplikacije) ili audio (kada je potrebno sinhronizovati usne na prirodni govor).



Sl.3. Faze učenja i sinteze u generisanju virtuelnog govornika [4].

U ovom radu se razmatra audio kao ulaz. Kasnije će biti prikazano da ako se jedinica za sekvenciranje

digitalizovanog signala govora zameni sa *text-to-speech* (TTS) jedinicom, rešenje se proširuje i na slučaj kada je ulaz tekst.

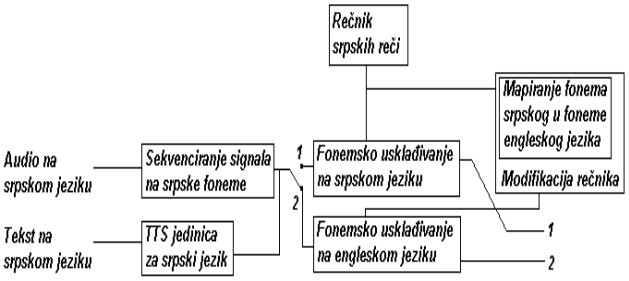
Fonemsko usklađivanje (*phonetic alignment*) se odnosi na vreme trajanja samih fonema, kao i na vreme trajanja između njih. Problem sa usklađivanjem imaju i vizemi (*visemic alignment*) pošto se obavi mapiranje fonema u vizeme. Za dobijanje odgovarajućih rezultata neophodno je izvršiti intenzivna treniranja velike datoteke standardnih rečenica. [5]

IV. Mapiranje fonema i vizema

A. Mapiranje fonema

Foneme engleskog i srpskog jezika možemo prikazati u skladu sa IPA (*International Phonetic Alphabet*) standardom. Engleski nema, kao srpski jezik, odgovarajući pisani simbol za svaki fonem u svom jeziku. U engleskom jeziku ima 26 slova, a broj fonema se kreće od 44, pa čak do 55. Srpski, sa druge strane, ima 30 slova i 30 fonema.

Postoje dve mogućnosti na ulazu u sistem. Pošto je dostavljen audio na srpskom jeziku, potrebno je razviti modul za prepoznavanje srpskih fonema. Prva mogućnost je da jezik, koji se koristi za treniranje ovog sistema za prepoznavanje, bude jednak jeziku u kome je dostavljen ulazni audio (srpski jezik). U ovom slučaju nema većih problema ako su sve jedinice za ovakav sistem na raspolaganju. Druga mogućnost je da ako na raspolaganju već imamo sistem za prepoznavanje fonema u engleskom jeziku, onda je moguće upravo njega iskoristiti i za srpski jezik. Jednom kada se fonemsko usklađivanje postigne, teško da izlazi imaju bilo kakvu zavisnost od jezika koji je korišćen. To se može iskoristiti u sintezi virtuelog govornika na proizvoljnom jeziku. Na sl.4. prikazane su obe mogućnosti, označene brojevima 1 i 2.



Sl.4. Blok šema sistema za prepoznavanje fonema.

Ako se koristi druga mogućnost, neophodan je blok koji obavlja mapiranje fonema srpskog u foneme engleskog jezika.

Postoje tri slučaja. Prvi slučaj je kada se reč srpskog jezika može predstaviti fonemima iz engleskog jezika. Kada se ona se ne može reprezentovati pomoću skupa fonema engleskog jezika, onda se ona mapira na način prikazan u sledećoj tabeli. Fonemi u engleskom jeziku koji se nikada ne javljaju u srpskom jeziku su suvišni i izostavljaju se. To su /θ/ i /ð/.

Mapiranje fonema se bazira na sličnim fonemima po zvučnosti. Na primer, ako nema tačnog fonema u engleskom jeziku koji odgovara srpskom fonemu, onda se bira fonem engleskog jezika koji je akustički sličan ili se bira string fonema engleskog jezika. U tabeli 2 sa ** označeni su akustički slični fonemi engleskog jezika, dok su sa * označena mesta gde je bio neophodan string od fonema engleskog jezika za aproksimaciju.

Dva akustički slična fonema različitih jezika mogu se mapirati u različite vizeme. Fonemsko mapiranje između dva jezika nije ,,1 na 1'', ali u odnosu na engleski jezik sigurno da postoje mnogo komplikovaniji jezici od srpskog [5], [6].

TABELA 2: MAPIRANJE FONEMA SRPSKOG JEZIKA.

p	p	***č*** /tʃʷ/	tS** /tʃ/	***e***	e /ɛ/
b	b	***dž*** /dʒʷ/	dZ** /dʒ/	***i***	I /i/
m	m	***š***	S /ʃ/	***o***	O /ɔ/
f	f	***ž***	Z /ʒ/	***u***	U /u/
v /ʋ/	v** /v/	***s***	s	***c*** /ts/	ts*
t	t	***z***	z	***đ*** / ***dz̟*** /	dj*
d	d	***n***	n	***ć*** /tɕ/	tj*
k	k	***l***	l	***j***	j
g	g	***r*** /r/	r** /ɹ/	***lj*** /ʎ/	lj*
h /x/	h** /h/	***a***	A /ɑ/	***nj*** /ɲ/	nj*

B. Mapiranje vizema

Za prepoznavanje vizema prikazana su četiri rešenja. Jedno rešenje je mapiranje odgovarajućih fonema iz srpskog jezika na vizeme engleskog, čiju bazu vizema imamo na raspolaganju. U tabeli 3, prve tri kolone odgovaraju standardu MPEG-4 FA (za engleski jezik) [3].

U MPEG- 4 FA je definisano samo 14 statičkih vizema, koji se jasno među sobom razlikuju. Može se koristiti i podskup baze vizema definisane standardom, ali to izuzetno degradira kvalitet videa. U ovom slučaju se fonemi kojih nema u srpskom jeziku mapiraju vizuelno najbližim. To bi predstavljalo mapiranje srpskih fonema c, ć i đ u vizeme engleskih fonema z, tS(~č) i dZ(~dž).

Poboljšanje tog rešenja je dodavanje odgovarajućih vizema, kojih nema u engleskom jeziku (*) i brisanje onih kojih nema u srpskom jeziku (***), kao što je prikazano. Ovakvo mapiranje sadrži 16 statističkih vizema (do 17. vizema, bez 3.).

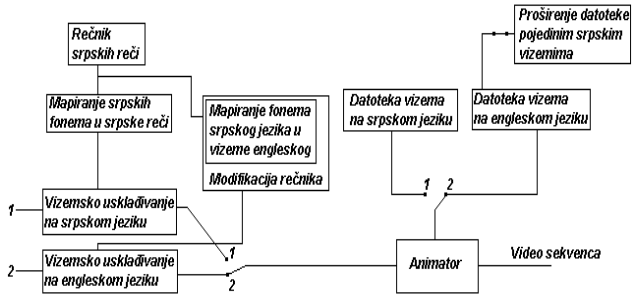
Ne postoji velika akustičkovizuelna razlika između fonema srpskog i engleskog jezika, pa možemo dozvoliti proširenje ovog skupa fonemom j (****) srpskog jezika, iako on postoji i u engleskom jeziku i ne pripada najužem skupu vizema definisanim standardom. To je učinjeno da bi se pokrili apsolutno svi fonemi srpskog jezika, a da pri tome nije narušena jednostavnost sistema. Ovo je ,, više na 1'' (*many-to-one*) mapiranje fonema u vizeme.

U tabeli 3 sa *, **, **** su označena kritična mesta, gde se nailazi na problem u skladu sa akustičnovizuelnim utiscima. Ta mesta predstavljaju aproksimacije pri mapiranju fonema srpskog jezika u vizeme engleskog.

TABELA 3: MAPIRANJE FONEMA SRPSKOG JEZIKA U VIZEME ENGLESKOG

Redni broj vizema	***Fonem (eng)***	***Primer (eng)***	***Fonem (srp)***	***Primer (srp)***
0.	none	-	nijedan	-
1.	p, b, m	Put, bed, mill	p, b, m	Pas, bol, med
2.	f, v**	Far, voice	f, v**	Far, voz
3.	T***, D***	Think, that	-	-
4.	t, d	Tip, doll	t, d	Tok, div
5.	k, g	Call, gas	k, g, h**	Konj, gas, hor
6.	tS, dZ, S	Chair, join, she	č**, dž**, š, ž**	Čas, džep, šah, žar
7.	s, z	Sir, zeal	s, z	Sat, zec
8.	n, l	Not, lot	n, l	Nos, lov
9.	r**	Red	r**	Rak
10.	A:**	Car	a**	Ako
11.	e	Bed	e	Evo
12.	I	Tip	i	Sit
13.	O	Top	o	Oko
14.	U	Book	u	Uvo
15.			c*	Car
16.			đ*, ć*	Đak, ćup
17.			j****, lj*, nj*	Ja, ljut, njuh

Najbolje rešenje je baza vizema srpskog jezika dobijena odgovarajućom opremom za Facial Motion Capture i intenzivnim treniranjem za svaki srpski fonem i odgovarajući vizem. Za sada su nam na raspolaganju prethodna tri rešenja. Sve ove mogućnosti u prepoznavanju vizema prikazane su na sledećoj slici u okviru blok šeme.



Sl.4. Blok šema sistema za prepoznavanje vizema.

Sam izraz lica nije strogo vezan za konkretan fonem, već i za prethodni i sledeći fonem, tako da je neophodno uvesti bidirekcionu predikciju. Obično se koristi *forward* predikcija, odnosno sledeći frejm koristi prethodni. Kod *backward* predikcije, odnosno korišćenja narednog frejma, povećava se kašnjenje u video sekvenci i zauzima se veći memorijski prostor na strani dekodera. U MPEG-4 FA prelazak iz jednog vizema u sledeći je definisan pomoću spajanja dva vizema u skladu sa težinskim faktorima. Još uvek nije jasno kako se standard može koristiti i u visokokvalitetnim animacijama vizuelnog govora.

V. ZAKLJUČAK

Virtuelni lik, koji stvara iluziju da govori na srpskom jeziku, imao bi veliku primenu, naročito u oblasti advertising-a i web i mobilnim aplikacijama. Ovde je pokazano kako je moguće iskoristiti gotove sisteme u engleskom jeziku za srpski jezik pomoću mapiranja fonema i vizema. Označena su mesta koja zadaju najveće probleme pri aproksimativnom mapiranju.

Veliki broj naših programera radi u oblasti MC-a, tako da je za očekivati da će u skorijoj budućnosti biti moguće koristiti skupu MC opremu i u našoj zemlji radi smanjenja troškova produkcije. Time će biti omogućeno pribavljanje najpreciznijih podataka o licu koje izgovara reči srpskog jezika. Princip „jedan glas – jedno slovo'' dodatno olakšava razvijanje jedinice za prepoznavanje fonema srpskog jezika i odgovarajućih sistema za treniranje.

LITERATURA

[1] M. Gleicher, "Animation From Observation: Motion Capture and Motion Editing", University of Wisconsin, Madison, Computer Graphics 33(4), p51-54, March 2000.

[2] A. Askovic, *Motion capture u našim uslovima*, Svet kompjutera, Beograd, februar 2005.
Available: http://www.sk.co.yu/2005/02/skpr02.html

[3] I. S. Pandzic and R. Forchheimer, "MPEG-4 Facial Animation - The Standard, Implementation and Applications", Link¨oping University, Sweden, 2002.

[4] Z. Deng, "*Data-driven facial animation synthesis by learning from Facial Motion Capture data*", University of Southern California, In Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy (Computer science), May 2006.

[5] T. A. Faruquie, C. Neti, N. Rajput, L. V. Subramaniam, A. Verma, "*Translingual visual speech syntehesis*", IBM India Research Lab, New Delhi 110016, India, + IBM T. J. Watson Research Center, Yorktown, Heights, NY 10598, USA, February 2000.

[6] C. Engström, Examensarbete i Talkommunikation, "*Articulatory Analysis of Swedish Visems*", Institutionen för tal, musik och hörsel Kungliga Tekniska Högskolan, 100 44 Stockholm, September 2003.

ABSTRACT

The paper considers the possibilites of using already trained systems in English for generating virtual speakers in Serbian according to the MPEG-4 FA standard. Mapping phonems and visems performed as one of the easiest ways for establishing a talking face in Serbian. This can especially be used in advertising and web and mobile applications. Facial Motion Capture cannot be neglected as a most precise method for gathering facial data. A great deal of programmers from underdeveloped countries is working in this area. Moreover, having a corresponding symbol for each symbol makes recognition and training systems in Serbian much easier to develop.

PHONETIC AND VISEMIC MAPPING FOR VIRTUAL SPEAKER IN SERBIAN

A. Gavrovska