Poređenje emocija u arapskom prirodnom i sintetizovanom govoru

Mohammad Al-Abbushi

***Sadržaj* — Cilj ovog rada je da se uporedi snimljeni glas govornika koji je pročitao određene rečenice na arapskom jeziku, sa i bez emocija, i da se uporedi sa sintetizovanim govorom pomoću TTS softvera za arapski, kao i sa ilustracijama istih rečenica i emocija datih u literaturi.**

***Ključne reči* — Prozodijski parametri: intenzitet, pitch, F0 i trajanje, slike snimljenog glasa.**

I. UVOD

POREĐENJE ljudskog glasa sa sintetizovanim govorom nije uobičajeno i generalno je teže razumljivo [1]. Poznati su nedostaci računarskih sistema za konverzaciju teksta u govor, odnosno mašina za čitanje. Uloga emocija u govoru je da se obezbedi intonacija govora tako da se može tumačiti namera govornika, a ovo je bitno i u sintetizovanom govoru. TTS sistemi za sintetizovanje govora iz teksta vrše simulaciju emocija, ako su tako projektovani. Postoje dva ugla da se sagledaju emocije: (1) Generativni model (govornik) koji zavisi od mentalnog i fizičkog stanja govornika, sintakse i semantike izgovora, i (2) Akustični model (slušalac) koji opisuje parametre akustičkih signala sa strane slušaoca [1], [2].

U radu će u II poglavlju biti predstavljeni moduli arapskog TTS sistema sa čijim sintetizovanim govorom je vršeno poređenje ljudskog glasa. U III poglavlju date su ilustracije primera prirodnog i sintetizovanog govora čije je poređenje vršeno. Uočeni su i analizirani osnovni prozodijski parametri i izvedeni određeni zaključci o njihovom prisustvu u sintetizovanom u odnosu na ljudski govor.

II. ARAPSKI TTS SISTEM

Sistem za pretvaranje teksta u govor kao ulaz prihvata pisani tekst, a kao izlaz generiše sintetizovani govor. Ovakvi TTS sistemi imaju veliku važnost u mnogim primenama, kao što su mašine za čitanje za slepe, prijem elektronske pošte telefonom, pristup velikim bazama podataka telefonom i u drugim primenama, [3], [4].

A. Opšti opis sistema:

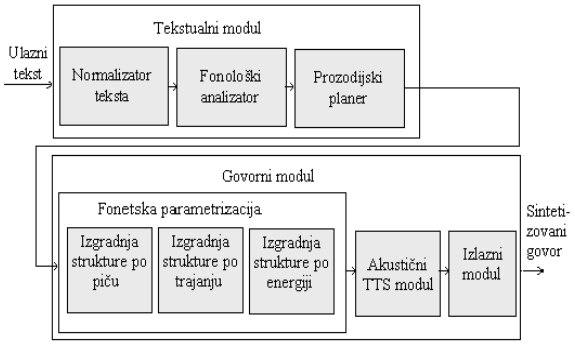
Na slici 1 je prikazano opšta arhitektutura sistema, koji se sastoji od tri osnovne komponente: tekstualnog modula, prozodijskog modula i izlaznog modula. Tekstualni modul uključuje normalizator teksta, fonološki analizator i prozodijski planer.

Mohammad Al-Abbushi, student Fakulteta tehničkih nauka u Novom Sadu, Srbija (e-mail: abushiali@yahoo.com).

Analizator teksta obavlja neke standardne zadatke u normalizaciji teksta kao što su pretvaranje cifara u njihove ekvivalente u rečima i zamena nekih standardnih skraćenica celim rečima.

Fonološki analizator za arapski jezik je odgovoran za pretvaranje grafema u foneme (na bazi unapred zadatih pravila, uz opciju da se može aktivirati rečnik sa izuzecima), rastavljanje na slogove, kao i za dodeljivanje naglaska pojedinačnim slogovima.

Prozodijski planer obavlja izvesno prozodijsko planiranje na nivou teksta.



Slika 1**:** Opšta arhitektura sistema

B. Pretvaranje reči u slogove:

Slog je segment govora koji sadrži samoglasnik koji predstavlja jezgro sloga i suglasnike koji ga okružuju. Prednost ovako duge jedinice govora je u tome da u sebe uključuje ko-artikulacione efekte između susednih segmenata. To je razlog što se kao jedinice za sintezu koriste slogovi, [5]. Pretvarenje arapskih slova u zvuk je jednostavno preslikavanje 1 na 1 između ortografske i fonetske transkripcije, za datu akcentaciju. Na slici 2 prikazana su mesta i način izgovora arapskih suglasnika u sistemu, [6].

C. Prozodija govora:

Prozodijske karakteristike su one koje više karakterišu iskaz kao celinu, nego što imaju lokalni uticaj na individualne segmente zvuka. U računarskom govoru, iskaz se sastoji od rečenica iako u prirodnom govoru iskaz može biti mnogo duži. Te prozodijske jedinice su međusobno čvrsto povezane. Promene u dinamici glasa se dešavaju u tri dimenzije: pič, trajanje i amplituda [7]. Ove dimenzije su u prirodnom govoru neraskidivo povezane.

MANNER OF ARTICULATION				PLACE OF ARTICULATION										
				Bilabial	Labio-Dental	Inter-Dental	Alveo-Dental Alveolar	Palatal	Velar	Uvelar	Pharangeal	Glottal
	Stop (Plosive)	Voiced	Pharyngealized	ب b			د d ض D	ج j		ق q		
		Unvoiced	Pharyngealized				ت t ط T		ك k			ء ʔ
	Fricative	Voiced	Pharyngealized			ظ Đ	ز z ذ Đ			غ G	ع ʕ	
		Unvoiced	Pharyngealized		ف f	ث T	س s ش S	ي j	خ x		ح ħ	ه h
	Affricate	Voiced					ن N					
	Nasal	Voiced		م m			ن n					
	Trill	Voiced					ر r					
	Lateral	Voiced					ل l					
	Semivowel	Voiced		و w				ي j				

*U nekim retkim slučajevima mogu delovati kao grleni.

Slika 2: Lista arapskih suglasnika

III. PRAVILA ZA EKSTRACIJU ZA RAZLIČITE EMOCIJE

D. Metodologija:

Najvažniji akustični parametri koje treba razmatrati prilikom sinteze emocija su prozodijski parametri: visina tona, trajanje i intenzitet [1], [2]. Varijacije svakog od ovih parametara su opisane preko sledećih sub-parametara [8], [1], [9], [10]:

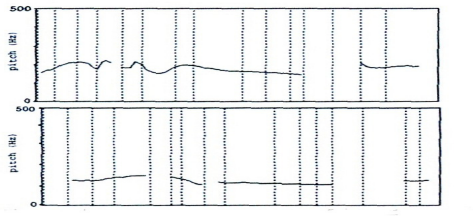
Parametar F0: opseg F0 (razlika između $F0_{max}$ i $F0_{min}$), promenljivost (stepen promenljivosti: visok, nizak,...), prosek F0, nagib konture (oblik nagiba konture), džiter (neregularnosti između uzastopnih glotalnih impulsa) i promena visine tona u skladu sa klasom fonema.

Parametar trajanja: brzina promene tokom govora i pauze, promena trajanja u skladu sa klasom fonema i promena trajanja u skladu sa visinom tona (pič).

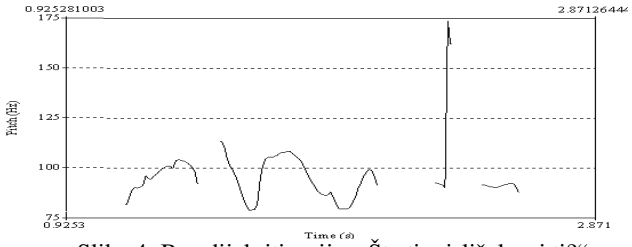
Parametar intenziteta: promena intenziteta.

E. Snimanje i analiza:

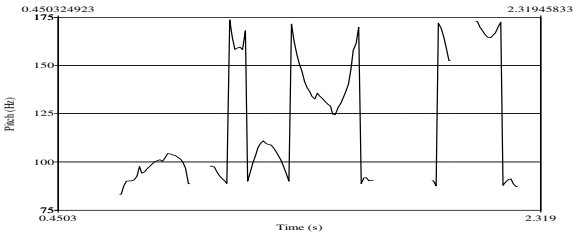
Za svaku emociju odabrano je po pet rečenica. Svaka rečenica je snimljena po dva puta, jednom bez emocija, a drugi put sa unapred odabranom emocijom. Sve ove rečenice analizirane su pomoću PRAAT sistema da bi se utvrdili prozodijski parametri. Na sl. 3. date su konture F0 za emociju besa i bez nje, kao što je dato u literaturi. Na slikama 4. i 5. ista rečenica je snimljena pomoću arapskog TTS sintetizatora, bez i sa dijakritizacijama. Dijakrizacija daje jednosmislenost značenja reči, oslanja se na proširenje osnovnog istraživanja u oblasti prirodnog izvršavanja jezika. Određuje značenje svaki reči, kao i način na koji će ona biti izgovorena, (tj. poseban govorni znak koji se piše iznad ili ispod slova) koji zamenjuje većinu samoglasnika, a to su Fatha, Kasra, Damma i Sukun (ـَ , ـِ , ـُ , ـْ), samo druga crta je ispod slova, a ostale su iznad.



Slika 3: Emocija besa bez i nje:
„Šta ti misliš, ko si ti?“,"من تظن نفسك"

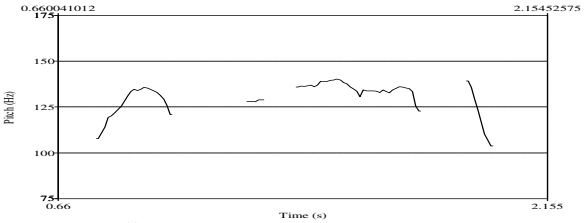


Slika 4: Bez dijakritizacije: „Šta ti misliš, ko si ti?“

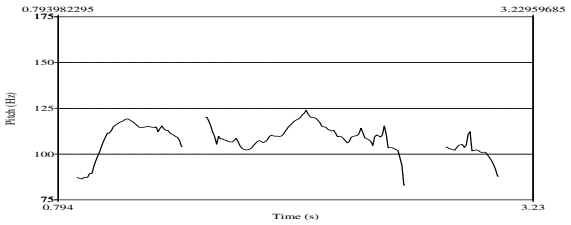


Slika 5: Sa dijakritizacijom: „Šta ti misliš, ko si ti?“

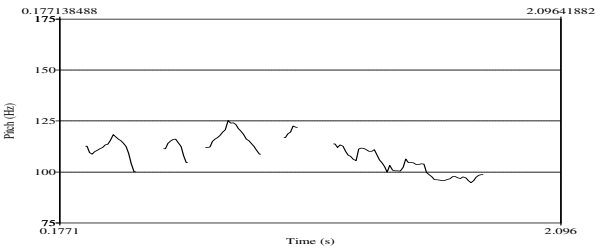
Na slikama 6-15 su prikazani snimljeni glasovi rečenica sa emocijama i bez njih, za iste rečenice koje je izgovorio čovek. Slike od 6 do 10 su bez emocija.



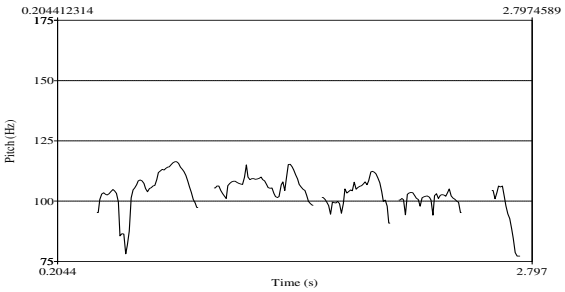
Slika 6: „Šta ti misliš, ko si ti?“ "من تظن نفسك" (neutralan)



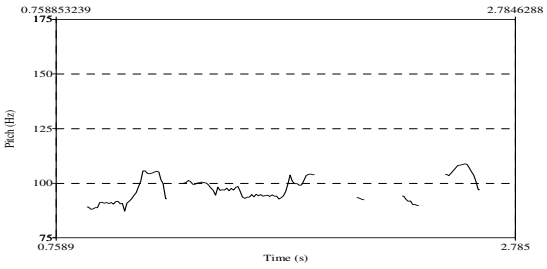
Slika 7:“Nema više oblaka na nebu“ ,"زالت الغيوم من السماء" (neutralan)



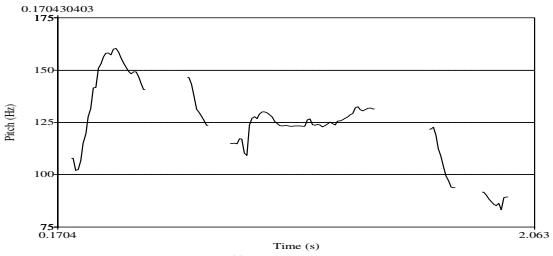
Slika 8: „Danas sam tako tužan!“ “انا حزين جدا اليوم!" (neutralan)



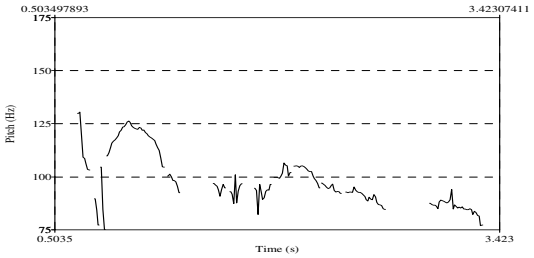
Slika 9: „O, Bože! Kakav zastrašujući prizor!“ "يا إلهي ما هذا المنظر المخيف" (neutralan)



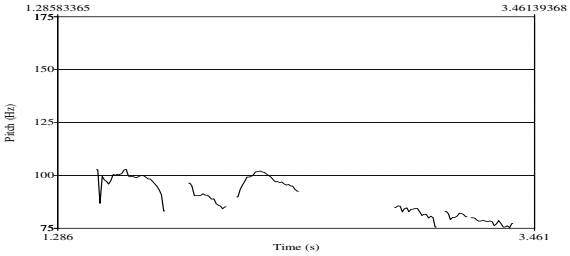
Slika 10: „Kakav divan prizor!“ ,"يا له من منظر جميل" (neutralan)



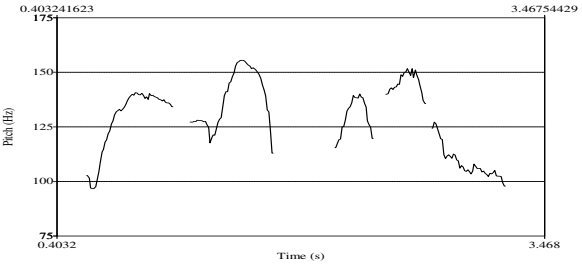
Slika 11: „Šta ti misliš, ko si ti?“ (bes)



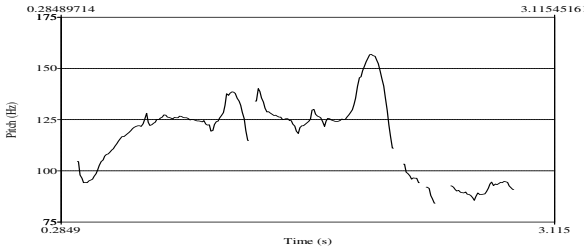
Slika 12**:** “Nema više oblaka na nebu“ (radost)



Slika 13: „Danas sam tako tužan!“(tuga)



Slika 14: „O, Bože! Kakav zastrašujući prizor!“ (strah)



Slika 15: „Kakav divan prizor!“ (iznenađenje)

IV. ANALIZA REZULTATA

U ovom poglavlju uočeni su i analizirani osnovni prozodijski parametri i izvedeni određeni zaključci o njihovom prisustvu u sintetizovanom u odnosu na ljudski govor.Kao što je u poglavlju III.A definisano u skladu sa literaturom najvažniji prozodijski parametri za poređenje i analizu prisustva emocija kako u prirodnom, tako i u sintetizovanom govoru su parametri: (1) F0, (2) trajanje i (3) intenzitet. Rezultati poređenja ovih parametara kod prirodnog i sintetizovanog govora, sa i bez emocija, sažeti su u Tabeli 1.

Rečenica	**Prozodijski Parametri**			
		Ljudski govor		**Sintetizatovan govor**
Šta ti misliš, ko si ti?		neutralan	bes	bez dijak.	sa dijak.
	Prosečan F0:	131.2 Hz	125.0 Hz	95 Hz	121 Hz
	Opseg F0:	33.8 Hz	77.5 Hz	90.5 Hz	104.5Hz
	Nagib F0:	118 Hz/s	147 Hz/s	253 Hz/s	517 Hz/s
	Džiter F0:	0.010	0.012	0.020	0.021
	Brzina govora:	2.6 reči/s	1.9 reči/s	2 reči/s	1.8reči/s
	Trajanje pauza:	173 ms	135 ms	nema	nema
	Intenzitet:	72 dB	84.4 dB	82.7 dB	83.34 dB
Nema više oblaka na nebu		neut.	radost	bez dijak.	sa dijak.
	Prosečan F0:	108 Hz	99 Hz	128.25Hz	126.14Hz
	Opseg F0:	48.8 Hz	63 Hz	107.1 Hz	107.5 Hz
	Nagib F0:	135 Hz/s	153 Hz/s	545 Hz/s	486 Hz/s
	Džiter F0:	0.011	143 ms	0.012	0.020
	Brzina govora:	2 reči/s	1.6reči/s	1.74reči/s	1.71reči/s
	Trajanje pauza:	104 ms	143 ms	nema	nema
	Intenzitet:	68.5dB	82dB	83.51 dB	83.33 dB

		neut.	tuga	bez dijak.	sa dijak.
Danas sam tako tužan!	Prosečan F0:	109Hz	90 Hz	126.9Hz	125.21Hz
	Opseg F0:	30Hz	27.8 Hz	107.1Hz	106.13Hz
	Nagib F0:	104Hz/s	74.5 Hz/s	500 Hz/s	485Hz/s
	Džiter F0:	0.010	0.01	0.020	0.012
	Br. govora:	2.3reči/s	1.8reči/s	2 reči/s	1.54 reči/s
	Traj.pauza:	nema	nema	nema	nema
	Intenzitet:	71 dB	68.1 dB	83.32 dB	83.7 dB
		neut.	strah	bez dijak.	sa dijak.
O, Bože! Kakav zastrašujući prizor!	Prosečan F0:	104 Hz	129 Hz	130 Hz	129.2 Hz
	Opseg F0:	40 Hz	58.5 Hz	107.21Hz	106.7 Hz
	Nagib F0:	186 Hz/s	72 Hz/s	487 Hz/s	461.1 Hz/s
	Džiter F0:	0.020	0.010	0.014	0.017
	Br. govora:	2.6reči/s	1.6 reči/s	2 reči/s	1.9 reči/s
	Traj. pauza:	nema	nema	nema	nema
	Intenzitet:	73 dB	63.5 dB	82.6 dB	82.81 dB
		neut.	iznenađ.	bez dijak.	sa dijak.
Kakav divan prizor!	Prosečan F0:	97.4 Hz	117 Hz	132.81Hz	135 Hz
	Opseg F0:	21 Hz	72.7 Hz	107.4 Hz	107.6 Hz
	Nagib F0:	84.4 Hz/s	135 Hz/s	545 Hz/s	502.4 Hz/s
	Džiter F0:	0.010	0.010	0.017	0.020
	Br. govora:	2.6reči/s	1.9 reči/s	2.3 reči/s	2.3 reči/s
	Traj. pauza:	nema	nema	nema	nema
	Intenzitet:	68.3dB	64 dB	82.72 dB	82.56 dB

Kod prosečne vrednosti F0 za prirodan govor bez emocija i sintetizovan govor sa i bez dijakritizacije, postoji malo veća razlika između njih sa emocijama tuge i straha. Slično je i za emocije radosti sa i bez dijak., a veća razlika je za emocije iznenađenja, tuge i besa sa i bez emocija i bez dijak., a mala razlika za emociju besa sa i bez emocija sa dijak. Za emociju straha sa i bez dijak. skoro se ne menja. Za ospeg promena F0 prirodnog govora sa i bez emocije i sintetizovanog sa i bez dijak., postoje malo veća razlika između njih za emocije radosti i straha. Slično je i za emociju besa bez emocija i bez dijak., a veća razlika za emocije tuge sa i bez emocija sa i bez dijak., iznenađenje bez emocija sa i bez dijak. i besa bez emocija sa dijak. Za emocije besa i iznenađenja sa emocijom sa i bez dijak., postoje mala razlika. Brzina govora sa i bez emocija sa i bez dijak. ne menja se puno. Za prosečan intenzitet prirodnog govora sa i bez emocija i sintetizovan sa i bez dijak., postoje male razlike za sve emocije, sem kod besa i radosti. Ipak ne možemo izvući generalniji zaključak samo na osnovu jednog istraživanja. Treba više eksperime-nata da bismo postigli bolje rezultate i treba posvetiti više prostora za akustične modele koji opisuju parametre akustičkih signala sa strane slušaoca, da bismo dobili bolji kvalitet ljudskog glasa, a time bi imali kvalitetniji pič, jer se pič pokazao važniji od parametra trajanja. Svakako, postoji dosta prostora za istraživanja kako na pravi način uključiti emocije u automatski sintetizovani govor.

V. Zaključak

Na osnovu analize ilustracija prozodijskih parametara u datim primerima rečenica na arapskom jeziku koje su izgovorene sa i bez emocija, i koje su isto tako proizvedene mašinskim prevođenjem teksta u sintetizovani govor, zaključuje se da emocije nisu toliko izražene u sintetizovanom govoru kao u prirodnom ljudskom glasu, te da ima još dosta prostora za istraživanja kako na pravi način uključiti emocije u automatski sintetizovani govor. Ovaj rad je još jednom potvrdio da su za ocenu emocija najvažnija tri prozodijska parametra govora: F0, trajanje i intenzitet.

Zahvalnica

Ovaj rad predstavlja uvodna istraživanja za magistarski rad i nastao je pod rukovodstvom prof. dr Vlade Delića sa Fakulteta tehničkih nauka u Novom Sadu.

Literatura

[1] J. Cahn, "The generation of affect in synthesized speech" *Journal of the American Voice I/O Society*, Vol. 8. pp. 1-19, Jul. 1990.

[2] O. Pierre-Yves, "The production and recognition of emotions in speech: features and algorithms" *International Journal of Human-Computer Studies Journal of Human-Computer Studies*, vol.59 n.1-2, pp.157-183, Jul. 2003.

[3] Abdel-Gawad, K.H. ''Arabic Text-to-Speech System.'' *M.Sc. Thesis,* Cairo Univ., 1989.

[4] Abu-Elyazeed, M.F. ''An Arabic Text-to-Speech System.''*Ph.D. Thesis,* Cairo Univ., 1990.

[5] Anees, I."Arabic Sounds."Cairo: Dar El-Nahdah El-Arabeyah, 1961.

[6] SAMPA, *Speech Assessment Methods Phonetic Alphabet*, http://www.phon.ucl.ac.uk/home/sampa/home.htm.

[7] Haggard, M.P., Mbler, A. and Callow, M. ''Pitch as Voicing Cue.'' *JASA*, 47, (1970), 613-17.

[8] F. Zotter, "Emotional speech", at URL: http://spsc.inw.tugraz.at/courses/asp/ws03/talks/zotter.pdf

[9] I. R. Murray, M. D. Edgington, D. Campion, and J. Lynn, "Rule-based emotion synthesis using concatenated speech", *ISCA Workshop on Speech & Emotion*, N. Ireland, 2000, pp. 173-177.

[10] J.M. Montero, J. Gutiérrez-Arriola, J. Colás,E.Enríquez and J.M.Pardo, "Analysis and modelling of emotional speech in Spanich", at URL: http://lorien.die.upm.es/~juancho/conferences/0237.pdf

Abstract

The main goal of this paper is to investigate the possibilities to include emotions into synthesized speech. Audio-visual analzsis of several utterances with and without emotion are analised. Text to speech system for the Arabic Language is also used. Presented work incorporate five emotions in natural speech: anger, joy, sadness, fear and surprise, as well as an educational Arabic text-to-speech system.

EMOTION COMPARISON IN ARABIC NATURAL AND SYNTHESISED SPEECH

Mohammad Al-Abbushi