VIRTUALNI POGOVORNI AGENT EVA – UMETNA INTELIGENCA ZA BOLJ NARAVNO INTERAKCIJO Z NAPRAVAMI

IZIDOR MLAKAR, SIMONA MAJHENIČ, MATEJ ROJC
Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Maribor, Slovenija
izidor.mlakar@um.si, simona.majhenic@um.si, matej.rojc@um.si

Sinopsis Eden ključnih izzivov človeku podobnih vmesnikov je smiselna raba subtilnih nejezikovnih signalov v interakciji. Zato je to po eni strani povezano s sinhronizacijo, po drugi pa z zagotavljanjem 'pravilne' interpretacije. Nejezikovni elementi niso zgolj preprosti nadomestki jezikovne vsebine, ampak so dejanske (neintegrirane) sestavine govora. Zato je cilj prispevka razviti pogovorni model za ustvarjanje človeku podobnega pogovora in najti rešitev za čustveno ter personalizirano interakcijo med človekom in strojem. Model ponuja (i) platformo za ustvarjanje 'pogovornega' znanja in virov, (ii) okvir za načrtovanje in ustvarjanje objezikovnega obnašanja in (iii) okvir za izvedbo čustvenega in odzivnega objezikovnega obnašanja s pomočjo izražanja stališč, čustev ter gest, sinhroniziranih z govorom. V drugem poglavju osvetlimo trenutno stanje na področju utelešenih pogovornih agentov. V tretjem poglavju orišemo pogovorni model EVA, kjer je osrednja zamisel oblikovati različne oblike objezikovnega obnašanja (geste) v povezavi z neoznačenim besedilom in širšim družbenim ter pogovornim kontekstom. V četrtem poglavju opisujemo pridobivanje 'pogovornega' znanja in potrebnih virov, ustvarjenih s pomočjo označevanja spontanega dialoga in korpusno analizo. Skupaj s petim poglavjem nato opisujemo, kako te vire vključimo v dvostopenjski pristop samodejnega ustvarjanja objezikovnega obnašanja. Prispevek sklenemo s študijo primera in prikazom sinteze objezikovnega obnašanja utelešenega pogovornega agenta EVA (Rojc et al. 2017).

Ključne besede:
pogovorni agent s telesom,
nejezikovno obnašanje,
pogovorni model EVA,
sinteza pogovornega obnašanja,
personalizirana interakcija

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.11.2022.8
ISBN 978-961-286-675-4

THE EMBODIED CONVERSATIONAL AGENT EVA – ARTIFICIAL INTELLIGENCE FOR A MORE NATURAL INTERACTION WITH DEVICES

IZIDOR MLAKAR, SIMONA MAJHENIČ, MATEJ ROJC
University of Maribor, Faculty of Electrical Engineering and Computer Science, Maribor, Slovenia
izidor.mlakar@um.si, simona.majhenic@um.si, matej.rojc@um.si

Abstract One of the key challenges of humanoid interfaces is sensible usage of subtle non-linguistic signals in interactions. This is, on the one hand, connected with synchronisation, and on the other with ensuring the 'correct' interpretation. Non-linguistic elements are not merely simple substitutes for linguistic content, but actual (non-integrated) components of speech. The paper aims at developing a conversational model for creating humanlike conversations and at finding a solution for emotional and personalised humans-machine interactions. The model offers (i) a platform for creating 'conversational' knowledge and sources, (ii) a framework for designing and creating colinguistic behaviour, and (iii) a framework for the realization of emotional and responsive colinguistic behaviour. In the second chapter, we look at the state-of-the-art in the field of embodied conversational agents. In the third chapter, we outline the conversational model EVA. In the fourth chapter, we describe the acquisition of 'conversational' knowledge and the relevant sources. In the fifth chapter, we then describe how to incorporate these sources into a two-stage approach for the automatic creation of colinguistic behaviour. The paper is concluded with a case study and demonstration of the colinguistic behaviour synthesis with the embodied conversational agent EVA (Rojc et al. 2017).

Keywords: embodied conversational agen, non-verbal behaviour, conversational model EVA, conversational behaviour synthesis, personalized interaction

https://doi.org/10.18690/um.ff.11.2022.8 DOI
978-961-286-675-4 ISBN

University of Maribor Press

1 Uvod

Digitalni sistemi se vse bolj uporabljajo za naloge, ki jih običajno izvajajo ljudje. Napredki na področju vmesnikov govorjenega jezika, obdelave naravnega jezika in umetne inteligence so prispevali k vse večji dostopnosti in rabi pogovornih agentov (npr. virtualnih tutorjev, spremljevalcev in asistentov) – sistemov, ki posnemajo človeško interakcijo (Laranjo et al. 2018). Ob hitrem tehnološkem razvoju in vsesplošni digitalizaciji, bolj naravna in posledično bolj razumljiva interakcija z digitalnim vmesnikom predstavlja enega ključnih izzivov moderne interakcije. Človek namreč interakcijske cilje dosega skozi pogovor (Luger et al. 2016).

V zadnjih letih je ravno zato opaziti visoko raziskovalno aktivnost predvsem na področju pogovornih modelov, ki vključujejo animirane ali človeku podobne virtualne like, ki jih imenujemo virtualni pogovorni agenti (pogovorni agenti s telesom, angl. *Embodied conversational agents* - ECA) (Cassell et al. 2001). Bolj znani so zlasti sistemi, ki podpirajo govorni vmesnik kot, denimo, Applova Siri, Googlov Now, Microsoftova Cortana ali Amazonova Alexa (McTear et al. 2016). Novejše raziskave na področju interakcije človek-stroj kažejo, da je najbolj naraven in učinkovit način sintetične interakcije – tudi v visoko tveganih okoljih, kot je zdravstvo (Philip et al. 2020), skrb za duševno zdravje (Provoost et al. 2017), poučevanje (Kramer et al. 2020) in pomoč iz okolice pri samostojnem življenju (Queiros et al. 2018) – takšen, ki temelji na posnemanju naravnih modalnosti, denimo, sinhroniziran govor ter geste in mimika. Razumevanje, sprejemanje in zaupanje informacijam je namreč tesno povezano z nesemantičnimi signali (npr. čustvovanje in upravljanje diskurza), ki jih govorniki posredujejo s pomočjo vizualnih znakov in prozodije (Mlakar et al. 2019; Stal et al. 2020). Lahko bi celo rekli, da je temelj medosebne interakcije sinhrono vključevanje in povezovanje verbalnih z neverbalnimi kanali. Verbalni kanali nosijo simbolno/semantično interpretacijo sporočila z jezikovnimi in para-jezikovnimi značilnostmi interakcije, medtem ko neverbalni kanali služijo kot dirigent komunikacije (McNeill 2016: 4; Kopp in Bergmann 2017).

Jezikovni kanal torej s pomočjo jezikovnih in parajezikovnih elementov opisuje simbolično oz. semantično interpretacijo informacij, med tem ko nejezikovni kanal (npr. telesna govorica) govor organizira (McNeill 2016: 4). Nejezikovni kanal zajema koncepte, kot so prozodija, govorica telesa, čustvovanje ali sentiment. Ti koncepti

so večfunkcijski in delujejo na psihološki, sociološki in biološki ravni ter v vsakem časovnem okvirju (Church in Godin-Meadow 2017). Dejansko predstavljajo osnovo kognitivnih zmožnosti in razumevanja). Tako npr. nasmešek, pogosto s sočasnim smehom, igra pomembno vlogo pri gradnji povezave med udeleženci pogovora, še posebej vzpostavljanju družbenih vezi, ki ustvarjajo vljudno medosebno okolje (Esposito et al. 2015; Ochs et al. 2017). Še vedno pa raba teh subtilnih nejezikovnih signalov v interakciji predstavlja enega ključnih izzivov modernih vmesnikov. Problem izhaja iz sinhronizacije sinhronizacijo in iz zagotavljanja 'pravilne' interpretacije. Nejezikovni elementi namreč niso zgolj preprosti nadomestki jezikovne vsebine, ampak so dejanska sestavina govora. Pri govorjeni interakciji tako nejezikovno obnašanje prispeva več kot 50 odstotkov informacije, pomembne pri gradnji skupne osnove pogovora (Cassel et al. 2001). Še več, več kot 70 odstotkov socialnega pomena pogovora posredujemo z nejezikovnimi koncepti (Birdwhistell 2010).

Večina raziskovalcev se torej strinja, da so neverbalni elementi (tj. geste, mimika in čustva) bistvena sestavina interakcije. Da bi v uporabniku vzbudili odnos, se morajo verbalni in neverbalni elementi vključevati 'pravilno' in skladno s pričakovanji glede na njegove vhodne dražljaje (Ciechanowski et al. 2018). Če se jezikovni in nejezikovni komunikacijski kanali ne poravnajo pravilno, lahko ECA izvede gib brez pomena, ki ga dojemamo kot šum. Še več, predstavljen koncept lahko popači pomen ter z napačno poravnavo ustvari neprimeren družbeni kontekst (McKeown et al. 2015). S povečevanjem stopnje naravnanosti in modalnosti uporabniške izkušnje se bistveno povečujejo človekova pričakovanja in dojemljivost napak (Poria et al. 2017). Bolj, kot je odziv stroja podoben človeškemu, večji in močnejši bo negativni učinek, kadar pride do nesinhronosti (npr. manj naraven glas in manj naravna animacija). V tem oziru je samodejno tvorjenje pogovornega obnašanja še daleč od popolnosti ali naravnosti. Za zagotavljanje dobrih rezultatov je pogosto potrebno človeško posredovanje (Navarro-Cerdan et al. 2018). Konec interakcijskega cikla vedno predstavlja aktivni odziv uporabnika in ne signali ali sama interakcija; česar pa stroj še vedno ni zmožen 'razumeti' ali poustvariti (Opel in Rhodes 2018).

Cilj prispevka je predstaviti pogovorni model za ustvarjanje človeku podobnega pogovora in najti rešitev za čustveno in personalizirano interakcijo med človekom in strojem. Predstavljen model ponuja (i) platformo za ustvarjanje 'pogovornega' znanja in virov, (ii) okvir za načrtovanje in ustvarjanje neverbalnega obnašanja in (iii)

okvir za izvedbo čustvenega neverbalnega obnašanja s pomočjo izražanja stališč, čustev in gest, ki so sinhronizirane z govorom. V drugem poglavju bomo najprej predstavili koncept neverbalnega obnašanja oz. gest kot ključnega mehanizma za pripravo govora in ustvarjanja kohezivnosti interakcije. V tretjem poglavju orišemo pogovorni model EVA, kjer je osrednja zamisel oblikovati različne oblike neverbalnega obnašanja (geste) v povezavi z neoznačenim besedilom in širšim družbenim in pogovornim kontekstom. V četrtem poglavju opisujemo pridobivanje 'pogovornega' znanja in potrebnih virov, ki jih ustvarjamo s pomočjo označevanja spontanega dialoga in s korpusno analizo. Skupaj s petim poglavjem nato opisujemo, kako te vire vključimo v dvostopenjski pristop samodejnega tvorjenja obnašanja. Pričujoč pristop naslavlja (a) problem oblikovanja obnašanja (namen in načrtovanje obnašanja) in (b) problem izvedbe obnašanja (animacija z EVO). Prispevek sklenemo s študijo primera in prikazom sinteze neverbalnega obnašanja pogovornega agenta s telesom EVA (Rojc et al. 2017).

2 Neverbalno obnašanje in geste v interakciji

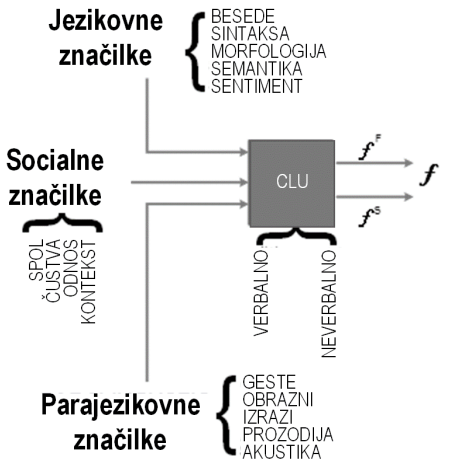
Neverbalni elementi interakcije predstavljajo pomemben del učinkovite komunikacije (Birdwhistell 2010; Trujillo et al. 2018). Vizualni vhodni/izhodni signali so večfunkcijski in delujejo na psihološki, sociološki in biološki ravni v vseh časovnih okvirih, npr. od trenutka do trenutka, ontogenetsko ter se razvijajo skozi čas glede na različna diskurzna okolja (Church in Goldin-Meadow 2017). Gestikuliranje in zmožnost izražanja informacij skozi neverbalne kanale postaja ključna metoda, s katero poosebiti in poenostaviti interakcijo med človekom in strojem. Pri klasični interakciji med dvema osebama so neverbalni signali, ki jih posredujemo skupaj z govorjeno vsebino ali pa tudi brez nje, ključni za kohezivnost diskurza. Lahko bi rekli, da jezikovni deli govorjenega jezika (to so besede, slovnica, skladnja) posredujejo simbolno/semantično interpretacijo sporočila, medtem ko neverbalni del (to so geste, izrazi, prozodija) nosijo družbeno komponento sporočila in so v vlogi dirigenta komunikacije. Posledično se neverbalni signali razlikujejo od splošnih motoričnih dejanj, saj predstavljajo to, kar je izraženo in prepoznano (npr. praskanje, ko razmišljamo ali mahanje v slovo). Allwood et al. (2005) tako raziskujejo povezavo med jezikovnimi in nejezikovnimi signali glede na funkcije komunikacije. Komunikacijo opredelijo kot vsoto glavnega sporočila in upravljanja komunikacije, ki pa jo ločijo na upravljanje interaktivne komunikacije (ICM) in upravljanje lastne komunikacije (OCM), obe pa se lahko izražata z jezikovno in nejezikovno

komunikacijo. Drugi raziskovalci (npr. Hoek et al. 2017; Chui et al. 2018; Lopez-Ozieblo 2018) se osredinjajo na semantiko. Eden najbolj zadevnih in široko rabljenih pristopov k večmodalnosti interakcije je Pierceova semiotična perspektiva (tj. 'pragmatika na dejanski strani') (Peirce 1965), ki proučuje pomen slik in povezanih vizualnih značilnosti pisnega besedila (Carroll et al. 2015, Queiroz in Aguiar 2015).

V nasprotju z omenjenimi pristopi pa semiotika proučuje tudi nejezikovne sisteme znakov. Pomen nejezikovnega obnašanja in pogovornih izrazov tolmači s proučevanjem za komunikacijo ključnih znakov in simbolov. Semiotika po Peirceu je trojiška in kot podskupine loči znake, ki so simboli; ikone in indekse. Vendar pa je njegova klasifikacija predvsem vezana na vizualni stimulus kot glavni nosilec interpretacije. Nasprotno Cooperrider (2017) ločuje med dvema sklopoma neverbalnega obnašanja (t. i. *gest*), in sicer med gestami ospredja in gestami ozadja. Geste ospredja vizualizirajo semantični del sporočila. Kot utemeljuje (Cooperrider 2017), jih rabimo zavedno in jih izvedemo z določeno mero truda. Pojavljajo se skupaj z govorom (npr. vizualizacija za razlago oz. oris) ali povsem brez hkratnega govora (npr. simboli). Zanje značilne kategorije so ikonske geste, zlasti tiste, ki izražajo prostorske odnose, ki lahko izražajo nujne, a jezikovno izpuščene informacije (Melinger in Levelt 2004). Geste ozadja pa so tiste, ki jih izvedemo z najmanj zavedanja oz. povsem podzavestno (Cooperrider 2017). Govorci navadno niso pozorni na njihove podrobne značilnosti in uporabo teh zelo hitro pozabijo. »So v ozadju govorčevega zavedanja, v ozadju poslušalčevega zavedanja in v ozadju interakcije« (Cooperrider 2017: 7). Kljub temu ločnica med tema dvema kategorijama ni tako jasna, saj je za nekatere tipe gest lahko zelo zabrisana (glej Cooperrider 2017: 193).

Neverbalni signali (ki vključujejo tudi in predvsem geste) dejansko omogočajo uporabnikom, da se na sintetični signal odzovejo naravno, začnejo izražati čustva in elemente, ki jih sicer vključujejo v medosebno interakcijo. Negativna plat pa seveda leži v dojemljivosti napak (pojav t. i. *uncanny valley*). Bolj kot je odziv naprave človeški, večja pričakovanja povzroči. Posledično bo odziv uporabnikov na pomanjkljivosti bistveno bolj negativen. Poglaviten vir morebitne neskladnosti (negativnega dojemanja) izhaja iz manka pogovornega znanja. Pogovorno znanje obsega razumevanje komunikativnih signalov, od jezikovnih, parajezikovnih do družbenih (Slika 1). Za razliko od jezika pa odnosov med signali ne moremo opisati z univerzalnimi pravili (kot je slovnica), ki bi ustvarili končen nabor vzrokov in

napovedali učinke. Da bi ustvarili delujoče, človeku podobne, odzive, moramo te signale spojiti v svoje področje človeku podobnih odzivov, za katere pa potrebujemo različne vire znanja. Kot prikazuje Slika 1, želimo oblikovati kompleksno funkcijo $\mathcal{f}$, ki je izražena kot fuzija informacije, ki jo lahko zajamemo skozi procesiranje naravnega jezika (angl. *Natural Language Processing* – NLP) in skozi procesiranje jezika telesa (angl. *Embodied Language Processing* – ELP). Obe domeni informacije pa skozi funkcijo $\mathcal{f}$ zlijemo v t. i. razumevanje pogovornega jezika (angl. *Conversaitonal Language Understanding* – CLU).



Slika 1: Razumevanje jezika kot model večmodalne fuzije
Vir: lasten.

Model predlagan na Sliki 1 temelji na konceptu 'večkanalne' predstavitve ideje, sočasno skozi osnovi avdio in video kanalov. Fuzija $\mathcal{f}$ se najprej oblikuje na kognitivni ravni s pomočjo simbolne fuzije ($\mathcal{f}^S$), kasneje na predstavitveni ravni s pomočjo fuzije oblike. Simbolno raven funkcije fuzije opredelimo kot

$$\mathcal{f}^S = f(L, P, S) \tag{1}$$

Tako vzpostavimo simbolično povezavo med različnimi signali iz različnih domen kognitivne lingvistike kot, denimo, sama lingvistika L, paralingvistika P in socialni kontekst S. Narava fuzijske funkcije f in korelacija z različnimi posameznimi signali in njihovimi prispevki je večinoma že zelo dobro opisana z novejšimi teorijami

kognitivne lingvistike in komunikativnega obnašanja. Vseeno pa ne moremo ozkega področja znanja kar 'zliti' v skupno strategijo. Da bi razumeli pomen $\mathcal{P}^S$ na simbolni ravni, predlagamo tolmačenje, ki je v skladu z McNeillovo (2008) teorijo skupne točke raste. To tolmačenje, izpostavljeno v Mlakar et al. (2019), izkorišča tako semiotiko po Peirceu (1965), nejezikovno obnašanje po Ekmanu in Friesenu (1971), kot kinezijo (Birdwhistell 2010; Maricchiolo et al. 2012). Skladno s teorijo upravljanja komunikacije (Guitella et al. 2009; Allwood 2014; McNeill et al. 2015) pa klasifikacija prav tako vključuje funkcije diskurza. Nejezikovni komunikacijski namen (NCI) v diskurzu zajema zgolj premikanje, ki služi nekemu komunikativnemu namenu (tj. prispeva k ustvarjanju pomena). Klasifikacija razlikuje med petimi različnimi vrstami takšnega gibanja, in sicer: ilustratorji, regulatorji ali adapterji, deiktiki ali kazalci, simboli ali emblemi in udarci.

Ilustratorji označujejo neverbalno obnašanje, s katerimi govorci ilustrirajo, kar govorijo. Sestavljeni so iz podskupine orisnih ilustratorjev, ki označujejo nejezikovno obnašanje, ki prikazujejo konkretno lastnost spremljajoče jezikovne vsebine, zaradi česar imajo v govoru jasno nanašalnico; podskupine ideografov, ki se nanašajo na konkretizacijo abstraktnega z določeno obliko; in prostorsko/dimenzijska podskupina, ki se nanaša na prostorske gibe, ki orišejo ali prikazujejo dimenzijska razmerja.

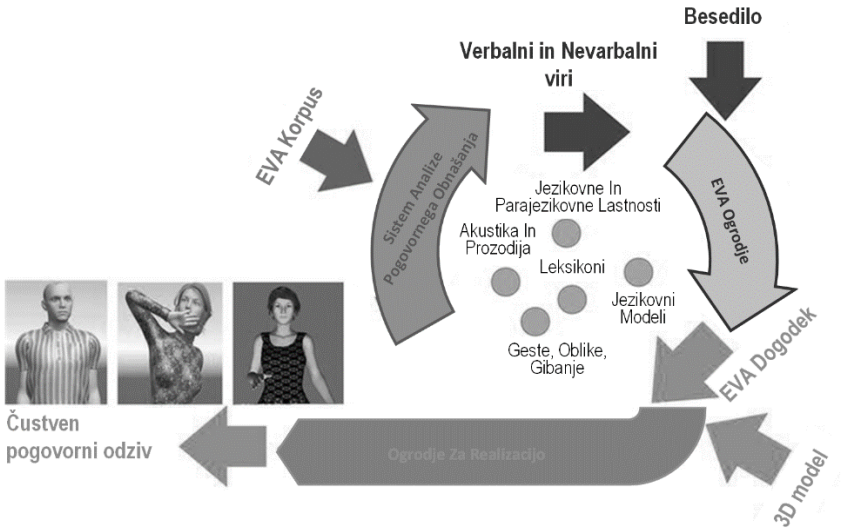
Regulatorji ali **adapterji** opredeljujejo nejezikovno obnašanje, ki lahko vsebuje nanašalnico v govorni vsebini ali strukturi. Primarno jih rabimo za prikaz metakonceptov, s katerimi upravljamo komunikacijo, izražamo mišljenje, napetost, negotovost ali druge družbene signale. Regulatorje še nadalje razdelimo na podskupino lastnih adapterjev, ki zajemajo obnašanje ob iskanju; podskupino regulatorjev komunikacije, ki vključujejo sekvenciranje in menjavo govornih vlog; podskupino regulatorjev afekta; podskupino manipulatorjev ter podskupino za družbene funkcije in norme.

Deiktiki se navezujejo na nejezikovno obnašanje, s katerimi se nanašamo na dejanske ali abstraktne zadeve (npr. predmete, kraje ali kazanje nazaj, kadar želimo prikazati preteklost). Četudi imajo dejansko nanašalnico v govoru, pa nejezikovno obnašanje ni nujno časovno povezano z njimi. Deiktiki zajemajo podskupino kazalcev; podskupino indeksov ali nanašalnih kazalcev in podskupino številčnikov. Skupina

NCI **simbolov** vključuje vse simbolne geste. Pogosto so kulturno specifične in imajo neposreden jezikovni prevod. **Udarci** so odrezani zamahi, ki dajejo ritem, ustvarjajo poudarke in s tem označujejo pomembnost kot tudi pritegnejo pozornost.

3 Pogovorni model EVA: Model za tvorjenje ekspresivnega sintetičnega obnašanja

Teoretičen model pogovornega prostora je orisan na Sliki 2. Model je zasnovan kot sredstvo, ki omogoča: (a) proučevanje narave naravnega obnašanje med sogovorniki (ljudmi); (b) ustvarjanje 'pogovornega' znanja v obliki lingvističnih, paralingvističnih jezikovnih in nejezikovnih značilk; (c) in preskušanje teorij skozi apliciranje znanja v različnih pogovornih situacijah.



Slika 2: model EVA: model za tvorjenje pogovornega obnašanja in oblikovanje čustvenih pogovornih odzivov na sintetičnih pogovornih agentih
Vir: lasten.

Model temelji na zamisli, da sta jezikovna in neverbalna poravnava ter sinhronizacija gonilni sili za afektivno in socialno interakcijo. *Sistem Analize Pogovornega Obnašanja* smo oblikovali za analizo prepletanja lingvističnih in parajezikovnih značilk z gestami, ki ga opazujemo v spontani interakciji med več govorci. Analiza temelji na video posnetkih večmodalnega korpusa EVA (Mlakar et al. 2019) in označevalni

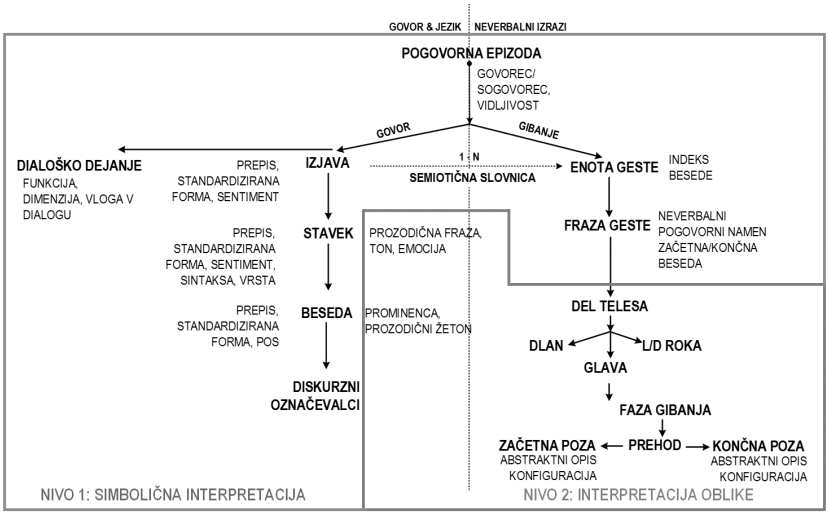
shemi EVA, razviti za opis kompleksnih odnosov neverbalnega obnašanja (predlagan v Mlakar et al. 2014). Korpus EVA je bil oblikovan za tvorjenje virov, ki jih potrebujemo za načrtovanje in ustvarjanje pogovornega obnašanja, tako jezikovnega dela (npr. sinteze od besedila v govor (angl. *Text-to-Speech* – TTS)) kot neverbalnih komponent (sinteza pogovornega obnašanja). Ključne vire predstavljajo leksikoni, jezikovni modeli, semiotična slovnica komunikativnega namena, leksikon pogovornih oblik, geste in gibi, akustične in prozodične značilnosti ter druge jezikovne in parajezikovne značilke (mdr. segmentacija na besedo/zlog, tip stavka, sentiment). Ključna zamisel ogrodja EVA (Rojc et al. 2014) je izkoristiti prej omenjene vire in načrtovati pogovorno obnašanje, ki lahko v interakciji izzove socialni/emocionalni odziv uporabnika. Rezultat ogrodja EVA prestavlja pogovorni niz, ki vsebuje sinhrono predstavitev informacije skozi govor in pogovorno obnašanje. Ker je načrtovano obnašanje že prilagojeno naravi in zmožnostim virtualnega agenta, EVA dogodek predstavlja direkten vhod za realizacijsko ogrodje EVA (Mlakar et al. 2018). To temelji na premisi, da je naravna večmodalna interakcija veliko več kot govor, ki ga spremljajo ponavljajoči se gibi okončin in obraza. Vloga tega ogrodja je animirati EVA dogodke s pomočjo 3D-modela. Realizacijsko ogrodje tako EVA dogodke preoblikuje v večmodalne, človeku podobne, pogovorne sekvence.

V naslednjih poglavjih bomo podrobneje predstavili tri glavne komponente pogovornega modela EVA, in sicer korpus EVA, ogrodje EVA ter realizator EVA.

4 EVA Korpus: Označevanje, segmentacija in kvantifikacija pogovora v pogovorne signale

Ključni cilj sheme na Sliki 3 je: i) na simbolni ravni identificirati pomene, ki jih je mogoče razbrati iz neverbalnih izrazov, kot funkcijo jezikovnih, parajezikovnih in socialnih signalov (npr. kdaj in kako gestikulirati) in ii) identificirati fizično naravo uporabljenih neverbalnih elementov (npr. kako izraziti), in sicer na ravni interpretacije nejezikovnih oblik. Koncept označevanje je tako dvonivojski. Prvi nivo na Sliki 3 imenujejo simbolična interpretacija. Uporabljamo ga za analizo interpretacije prepletanja različnih pogovornih signalov (npr. dialoška dejanja, geste, skladnja, diskurzni označevalci) in razumevanja, kako sodelujejo pri oblikovanju večmodalne predstavitve informacije. Simbolno označevanje omogoča identifikacijo

in podroben opis narave komunikativnih dejanj, ki se odvijajo med izmenjavo informacije. Označevanje oblike (oz. nivo interpretacije oblike) pa opisuje, kako izvesti neverbalne elemente, da pravilno "vizualizirajo" idejo/namen. Vizualizacijo dosežemo skozi prozodijo govora ter oblike in gibe, ki se pojavljajo ob govoru; npr. kako s premiki rok (leva in desna roka ter dlani), obraznimi izrazi, gibanjem glave in usmeritvijo pogleda poudarimo pomembne segmente, izražamo strinjanje/razumevanje ali sporočamo, da smo zaključili s podajanjem informacije in prosimo za odziv.



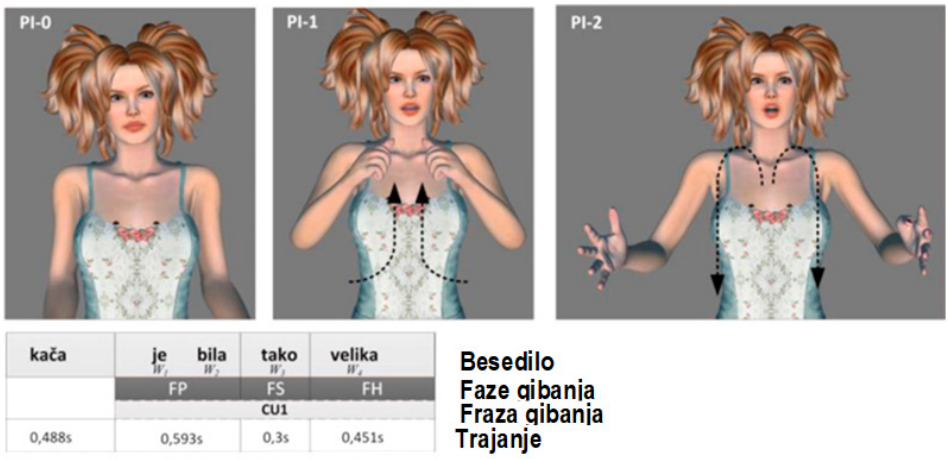
Slika 3: Topologija označevanja pogovornega obnašanja v EVA Korpusu: Verbalni in neverbalni kontekst pogovornih epizod.

Vir: lasten.

Simbolna interpretacija se tako ukvarja izključno z namenom neverbalnih komponent, ki ga klasificiramo z neverbalnim komunikacijskim namenom (NCI). Interpretacija oblik pa se ukvarja izključno z načinom izvedbe in vizualizacije. Da bi ju povezali, v predlaganem modelu uvedemo pojem semiotične slovnice. V njej je namen predstavljen kot razred/podrazred NCI. Vsak NCI pa zajema možne izvedbe neverbalnega obnašanja, s katerimi so govorci/poslušalci dosegli želeni namen. Prednost tega je, da je semantični prostor močno zmanjšan. Poleg tega pa je število eksplicitnih korelacij med besednimi sekvencami (besedami in frazami) in

motoričnimi spretnostmi zmanjšano na skupek razredov NCI nekaj podskupin. Gestikon predstavlja realizacijo koncepta semiotične slovnice.

Interpretacija oblike opisuje realizacijo pogovornega namena s prozodijo in telesnim gibom. Glavni cilj nivoja 2 je zagotoviti podroben opis, ki je blizu tako fizični realnosti (človeku) kot entiteti, ki jo realizira (npr. pogovorni agenti s telesom). V predlagani shemi so deli telesa ključni pri opazovanju in označevanju oblike. Pri tem sprejemamo zamisel utelešene kognicije, po kateri senzorično-motorične zmožnosti (zmožnost telesa, da se odzove na dražljaje z gibi), telo in okolje igrata pomembno vlogo pri razmišljanju. Označevalna shema EVA zato razlikuje med dlanmi, rokami, glavo in obrazom. Struktura oz. prozodija gibanja je nato opisana v obliki faz gibanja. Fazo gibanja, skladno s Kita et al. (1997), opišemo kot eno izmed petih, in sicer kot obvezna faza udarca ali kot opcijske pripravljalna faza, faza zadrževanja ter faza umika. Da bi dosegli izvedljivost 'giba' na dani entiteti, vsako izmed faz gibanja opišemo kot par začetne poze (P_S) in končne poze (P_E) ter pot T, po kateri je bil prehod med P_S in P_E izveden (točkasta puščica na Sliki 4) (Rojc et al. 2014). Pot T predstavimo skozi parametričen opis premika, ki vključuje zaporedje enostavnih vzorcev, kot so: linearno ali kot lok. Primer, kako uporabiti semiotično slovnico kot vir simbolične interpretacije in gestikon kot vir za realizacijo pogovornega namena, je podan na Sliki 4.



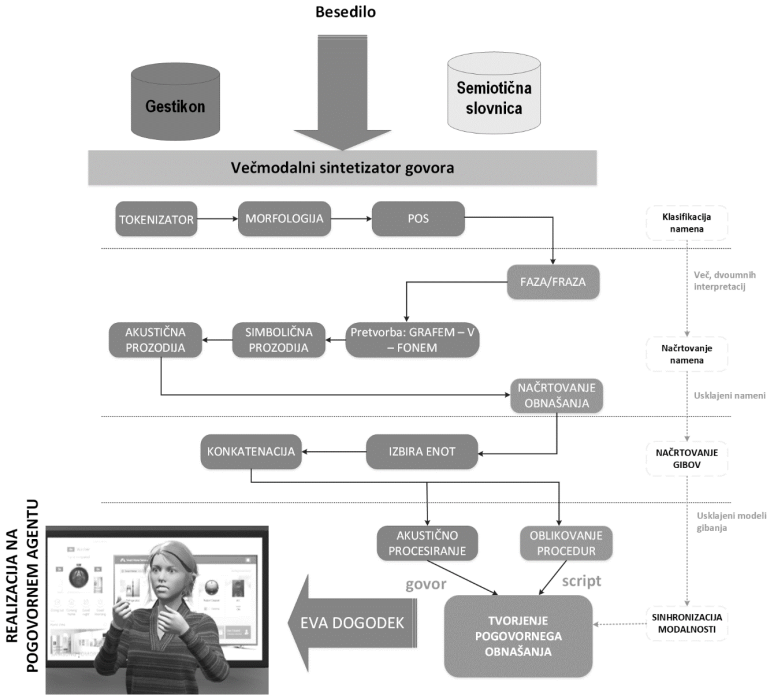
Slika 4: Vizualizacija pogovornega obnašanja na ECA ob uporabi semiotične slovnice in gestikona.
Vir: lasten.

Za 'vizualizacijo' stavka »Kača je bila tako velika« smo animirali eno od možnih interpretacij; serijo ikonsko metonimičnih gest, ki so rezultat sekvenc pomožnih glagolov ('je bila'), prislovov ('tako') in pridevnikov ('velika'). Kot pomensko jedro je bil prepoznan pridevnik 'velika'. Pripravljalna faza (F_P) je opredeljena z izvedbo premika med pozo PI-0 in PI-1 med besedama 'je' in 'bila'. Predvideno trajanje F_P je t= 593 ms. Vsebinsko najpomembnejša faza udarca (F_S) se bo izvedla ob prozodično najbolj poudarjeni besedi 'tako', njena oblika pa bo vizualizirala pomensko jedro, pridevnik 'velika'. Trajanje prehoda med PI-1 in P-2 je tako predvideno s časom t = 300 ms. Po izvedbi je algoritem, predstavljen na naslednjem poglavju, napovedal še fazo zadržanja, ki se izvede ob izgovarjavi pomensko jedro in traja 451 ms.

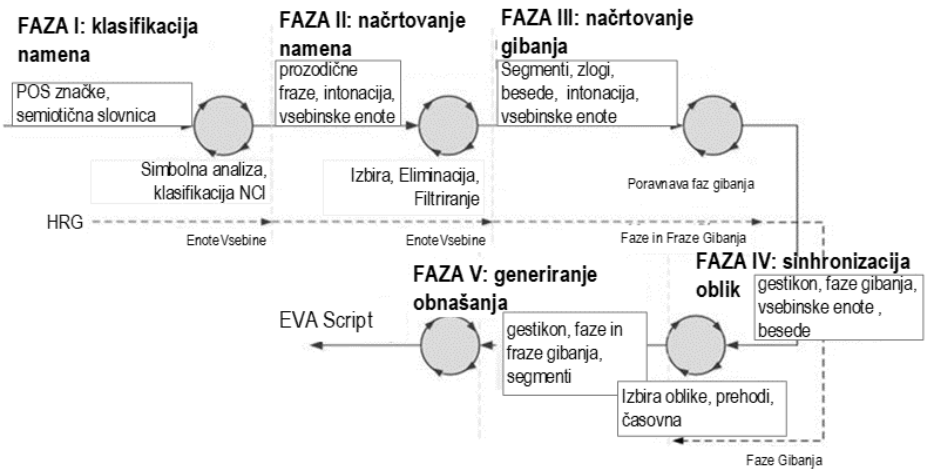
5 Ogrodje EVA: Generator pogovornega obnašanja

Algoritem načrtovanja pogovornih sekvenc (Slika 5) samodejno ustvari pogovorne sekvence, kot je sekvenca na Sliki 4, tako da za dano besedilo ustvari govor in neverbalne elemente. Njegova zasnova je podana na Sliki 5. Model temelji na štirih korakih: klasifikaciji namena, načrtovanju namena, načrtovanju gibov in sinhronizaciji modalnosti.

Proces sinteze govora, ki ga izvede algoritem na Sliki 5, pretvori splošno besedilo v pogovorni dogodek. Načrtovanje neverbalnih elementov je vključeno direktno v proces sinteze in lahko direktno izkorišča faze sinteze kot vir jezikovnih in prozodičnih značilnosti, ki so nujne za načrtovanje in sinhronizacijo neverbalnega obnašanja. Najprej algoritem izvede proces klasifikacije namena, ki identificira naravo govorne vsebine s pomočjo klasifikacije vzorcev besedila v semiotično slovnico. Pogovorni namen vhodnega besedila se opredeli v obliki klasifikacije glede na semiotični razredi. Rezultat prvega koraka je nabor možnih interpretacij vhodnega besedila. Proces načrtovanja namena nato vključuje izločanje interpretacije, ki ustreza prozodični strukturi govora. Za izbrani namen je nato iz Gestikona treba izbrati najustreznejšo interpretacijo. Postopek se izvede v koraku načrtovanja gibov, ki s pomočjo mehanizma cenilk izbere najustreznejšo izvedbo (t. i. model gibanja). Na koncu mora algoritem izvesti še časovno sinhronizacijo; tj. prilagoditi posamezne faze gibanja verbalnemu delu pogovorne sekvence. Slika 6 podrobneje orisuje postopek, ki je sestavljen iz petih faz.



Slika 5: Algoritem za načrtovanje in tvorjenje večmodalnih pogovornih sekvenc
Vir: lasten.



Slika 6: Algoritem za ustvarjanje čustvenega neverbalnega obnašanja.
Vir: lasten.

V **prvi fazi**, ki jo imenujemo klasifikacija namena, je vhod besednovrstno označeno (POS) besedilo in semiotična slovnica. Semiotična slovnica se uporablja za pripenjanje posameznih morfosintaktičnih sekvenc besedila na zadevne parametrične opise NCI. Algoritem išče najdaljše morfosintaktične sekvence, ki jih lahko najde v semiotični slovnici, pri čemer pa upošteva naslednji dve pravili:

Če je ob določeni besedi sekvenca x_A vrednost $x_A(S) \subseteq x_B(S)$, pri čemer spadata obe sekvenci k isti semiotični skupini, moramo sekvenco x_A zavreči.

Če je sekvenca x ob besedi j že zajeta v vsebinski enoti, ki se prične z besedo i in če ima enak pogovorni namen ($i < j$), jo zavržemo.

Vsebinska enota (CU) predstavlja parametrično interpretacijo sporočila v stavku/izreku. Stavki/izreki lahko vsebujejo več interpretacij in interpretacija CU se lahko delno ali povsem pokriva, s čimer pa prihaja do dvoumnosti in številnih neskladij. Posledično je treba v **prvi fazi,** pri načrtovanju namena, ta neskladja in dvoumnosti razrešiti s pomočjo integracije, eliminacije in izbiranja. Zato uporabljamo prozodične informacije (izstopanje, poudarek, prozodične fraze), kot to predvidevajo modeli TTS. Uporabljena prozodična informacija vključuje zloge z označenim naglasom, in sicer oznake PA, ki je najbolj izstopajoč, in NA, naglas besede. Vsebinske enote nato obravnavamo z naslednjimi pravili:

Vsaka vsebinska enota (CU) mora vključevati najbolj izstopajoč zlog (PA) znotraj dane prozodične fraze (B2 ali B3), razen pri naštevalnih primerih.

Vsak element CU se mora nahajati znotraj prozodične fraze (B2 ali B3).

Vsako prozodično frazo lahko predstavimo z največ enim konceptom neverbalnega gina, tj. ne več kot en element CU.

Kadar element CU vsebuje semiotični razred naštevanja, morajo meje CU ostati nespremenjene (mej prozodičnih fraz ne upoštevamo).

Element CU vključuje zlog PA, ki se mora nahajati znotraj mej prozodične fraze $B2{:}PA \in B2 \wedge PA \in CUi$.

V **tretji fazi**, ki jo imenujemo načrtovanje gibanja, opredelimo modele gibanja, s katerimi lahko vizualiziramo dani CU. Pri tem se ustvari model gibanja H, ki predstavlja animirano sekvenco oblik/položaja telesa, ki skupaj predstavlja neverbalni izraz. Za vsak H moramo določiti vsaj eno fazo udarca F_S, ki je poravnan z akustično prozodijo, kot jo definira TTS Za opredelitev faze udarca F_S smo uporabili naslednji pravili:

Faza udarca Fs se zmeraj nahaja ob PA besede in se konča skupaj s pripadajočim PA zloga.

Če beseda, ki predstavlja semiotični indikator I, za specifično CU ne vsebuje PA zloga, se v ta namen upošteva NA zloga.

V modelu gibanja H se zlogi, ki se pojavijo pred fazo udarca F_S, uporabljajo za pripravo faze gibanja F_P, segment 'sil', ki je tik pred prvim zlogom F_S, pa se lahko uporabi za fazo zadržanja gibanja F_H (zadržanje pred udarcem). Zlogi za Fs pa se lahko uporabijo za fazo umika F_R. V tem smislu strukturo obnašanja uporabljamo s pomočjo naslednjih pravil:

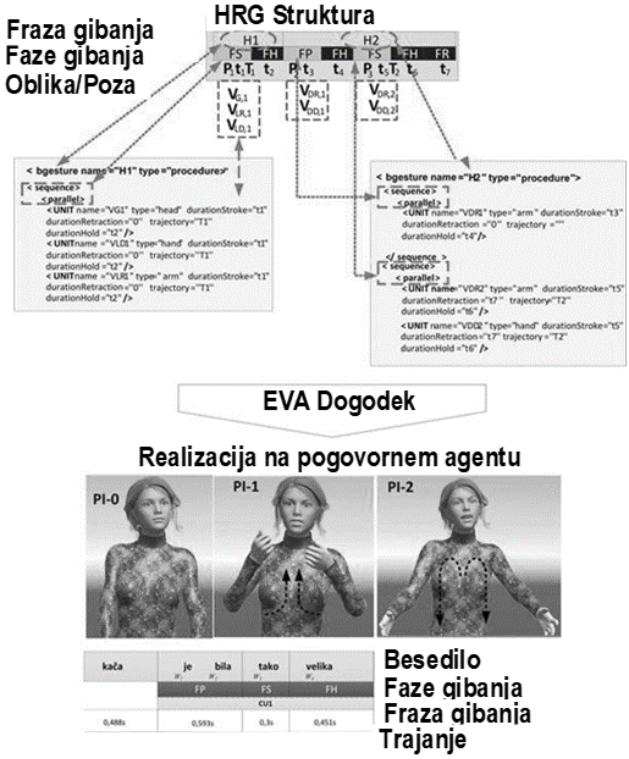
Faza pripravljanja Fp se začne pred fazo udarca Fs in traja od zloga NA do začetka faze udarca Fs.

Segment 'sil', ki lahko ima vrednost trajanja med besedami pri fazi pripravljanja Fp in fazi udarca Fs od nič in naprej, pa predstavlja t. i. fazo zadržanja pred udarcem (angl. hold before stroke), ki (če se pojavi) predstavlja pripravljeno idejo vsebine.

V **četrti fazi**, ki jo imenujemo sinhronizacija oblike, se gibanje časovno poravna s časovnimi značilkami jezikovne informacije (trajanje fonemov in premorov). Da bi določili najboljšo obliko V (ali položaj P) ter usmeritev T realizacije neverbalnega obnašanja, v *Gestikonu* izvedemo poizvedbo, ki temelji na morfosintaktičnih sekvencah, modelih gibanja in trajanja faz gibanja. Tako sprožimo iskanje možne konfiguracije oblike V znotraj F_S faz. Tako dobimo nabor možnih položajev telesa P za vsak F_S. Te položaje nato ocenjujemo s pomočjo funkcij primernosti (Rojc in Kačič 2011). Če v Gestikonu ni ujemanj, se izbere nabor najbolj primernih položajev v izbranem modelu CART (angl. classification and regression tree), pri čemer vsaki pripišemo najbolj primeren položaj P. Ko smo opredelili vse kandidate za položaj za vse predlagane F_S, se opredelijo še položaji za F_P, F_R in F_H, pri čemer za oblikovanje

prehoda med dvema položajema upoštevamo časovno strukturo, opredeljeno s trajanjem govornih enot, semiotičnega razreda, vrste faze gibanja, morfosintaktičnih oznak, prozodičnih značilk znotraj fraze.

V zadnji, **peti fazi**, ki smo jo poimenovali ustvarjanje neverbalnega obnašanja, predlagani model gibanja pretvorimo v proceduralni opis animacije. Vsaka faza gibanja se pretvori v simbolno, prozodično in prostorsko koherentno gibanje posameznega opazovanega telesa. Vsak model H opišemo v proceduralni sintaksi EVA-Script (Rojc et al. 2017) kot blok <*bgesture*>. Faze udarca F_S znotraj bloka <*bgesture*> predstavimo kot sekvence, bloke <*sequence*>. Fazi zadržanja F_H in faza umika F_R pa skozi atributa trajanja znotraj blok <*bgesture*> ali bloka <*sequence*>. Pretvorba gibanja modela H v dogodek EVA (neverbalno obnašanje, zapisano kot EVA-Script) je orisana na Sliki 7.

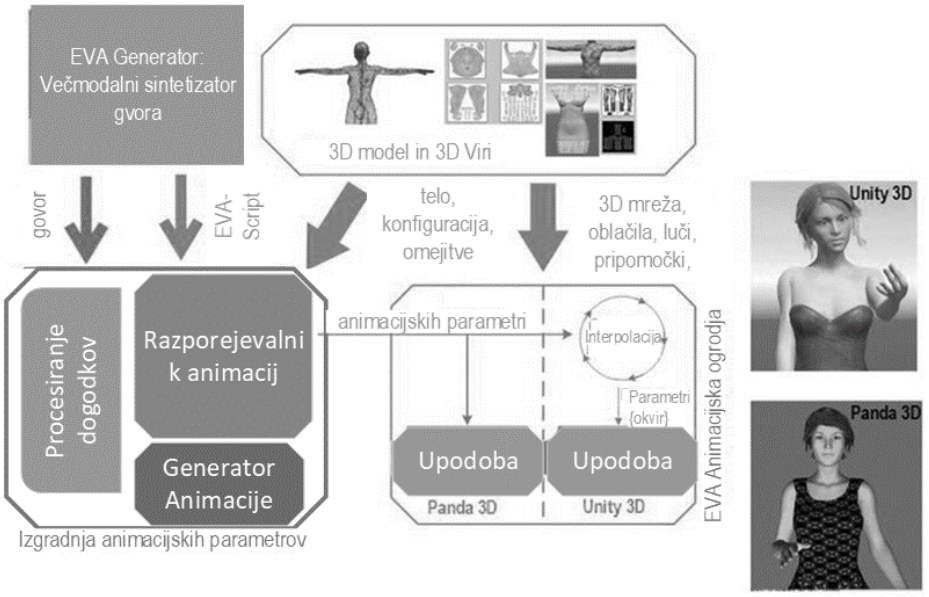


Slika 7: Realizacija stavka iz Slike 4 s pogovornim agentom EVA.
Vir: lasten.

Dejanska konfiguracija je opisana preko 3D-konfiguracije sklepov, ki je opisana kot element <*UNIT*> znotraj bloka <*sequence*>. Kadar so elementi <*UNIT*> združeni v bloku <*parallel*>, to označuje njihovo sočasno izvedbo. V nasprotnem primeru se konfiguracije oz. premiki iz enega 3D-sestava v drugi 3D-sestav izvedejo zaporedno. V naslednjem poglavju bomo predstavili realizator pogovornega obnašanja, s katerim proceduralni zapis v EVA-Script pretvorimo v animirano pogovorno sekvenco.

6 EVA Realizator pogovornega obnašanja: Animacija in vizualizacija govora na pogovornem agentu

Za animacijo načrtovanega obnašanja smo uporabili lastno razvito ogrodje realizacije – EVA (Mlakar et al. 2017. Ogrodje omogoča, da stroji z uporabnikom vzpostavijo bolj oseben stik, in sicer v obliki človeku podobne entitete, realizirane z večmodalnimi modeli interakcije, ki temeljijo na konceptu pogovora. Ogrodje je prikazano na Sliki 8:



Slika 8: Ogrodje za Realizacijo pogovornega obnašanja
Vir: lasten.

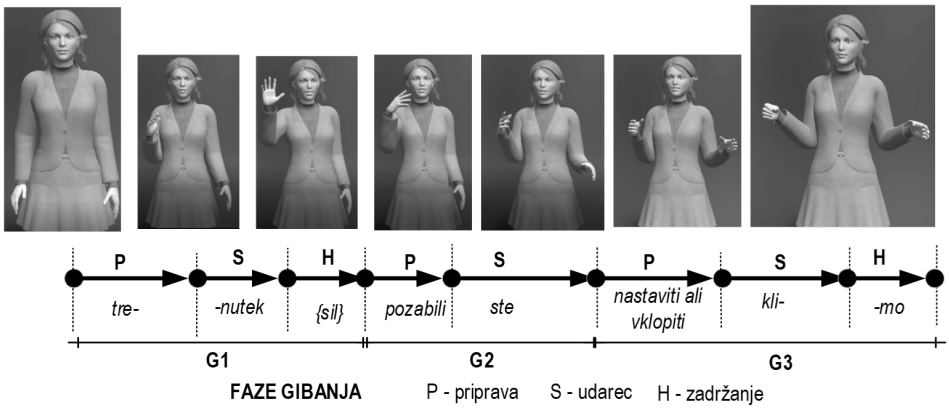
V grobem ogrodje sestavljajo komponente za gradnjo animacijskih parametrov, animacijska ogrodja za realizacijo animacije in 3D-viri, vključujoč mrežni model podobe agenta, oblačil in drugih elementov scene. Komponenta za gradnjo animacijskih parametrov se uporablja za pretvorbo neverbalnih dogodkov v nize animacijskih parametrov, ki jih animacijska ogrodja lahko pretvorijo v dejansko animacijo in prikažejo uporabniku. Da bi lahko formalni zapis obnašanja iz Slike 7 animirali, je treba EVA-Script zapis pretvoriti v nize animacijskih parametrov, ki jih je izbrano ogrodje za animacijo zmožno vizualizirati. To lahko dosežemo z uporabo neverbalnih značilk, opisanih v neverbalnih dogodkih pri 3D-virih, do katerih dostopa pretvornik. Komponente za gradnjo animacijskih parametrov prevede EVA-Script v animacijske parametre tako, da posamezne elemente v formalnem zapisu poveže z ustrezno kontrolno enoto agenta, vključujoč časovne (trajanje faze udarca, zadrževanja in umika) in prostorske značilnosti (končni položaj enote). Prevod se izvede z generatorjem animacije. Razporejevalnik animacij pa ustrezno interpretira sosledje v formalnem zapisu in pretvori v t. i. animacijski graf, ki ga realizira izbrano animacijsko ogrodje. Kot prikazuje na Sliki 8, predlagan model podpira dve animacijski ogrodji, in sicer Panda 3D in Unity 3D. Bistvena razlika med njima je v načinu izračuna interpolacije. Panda 3D interpolacijo med začetnim in končnim položajem izvede vnaprej. V Unity 3D pa se za izračun izvede na koncu vsakega video okvira. Izračun konfiguracije v naslednjem okviru daje bistveno večjo možnost glajenja animacije in reaktivnost obnašanja, saj lahko animacijo spremenimo ob vsakem koraku, celo med izvajanjem nekaterih korakov/sekvenc. Za gladek prehod razporejevalnik ne izvaja časovne prerazporeditve, ampak samo zamenja obstoječe segmente z novimi konfiguracijami. Posledično je virtualni lik bolj odziven in se lahko nemudoma odzove na spremembe v pogovornih, okolijskih in drugih kontekstih.

7 Demonstracija sinteze pogovornega obnašanja z modelom EVA

Podrobneje smo preučevali zaznano kakovost predstavitve informacij s posameznimi študijami primerov rabe. Primeri, ki sledijo, ponazarjajo praktične primere rabe:

Primer 1: Sinteza pogovornih izrazov z ECA EVA v okolju Pametni dom

ECA EVA: *»Tre**nutek**! Pozabili **ste** nastaviti ali vklopiti **kli**mo.«*



Slika 9: Sinteza pogovorne sekvence s pomočjo ECA EVA.
Vir: lasten.

Pri izreku zgoraj je model ustvarjanja obnašanja uporabil bazo znanja (tj. pravila) in vire, s katerimi je načrtoval in ustvaril opozorilno sekvenco v obliki naravnega jezika, ki vključuje sintetičen govor in tri prozodično poravnane pogovorne geste – G1, G2 in G3, kot je razvidno iz Slike 9. Vse tri geste vključujejo usmerjenost pogleda, gibanje leve in desne roke (ter dlani), pri čemer pa je desna roka prevladujoč del telesa. G1 prikazuje besedo 'trenutek' v kontekstu poudarka (tj. podrazred semiotičnega namena udarcev, IB). Faza udarca se zgodi na lokalno najbolj izstopajočem delu govorne sekvence (tj. 'trenutek'). Oblika dlani na koncu udarca je ena od značilnih oblik v Korpusu EVA, ki se pojavlja skupaj z udarci, ki pa so povezani s konteksti kot čakanje, zadrževanje ali ustavitev. Ker govorjeni vsebini sledi kratka tišina ({sil}) in je izrek vzkličen, se je model odločil tišino prikazati s fazo zadrževanja gibanja. G2 predstavlja tipičen referenčen izrek (tj. podskupina semiotičnega namena nanašalnih deiktikov, Dr), ki se prej nanaša na sogovornika kot pa na tretjo osebo ali predmet. Prozodično gledano je 'ste' najbolj izstopajoč del sekvence, kar je tudi razlog, zakaj se faza udarca pojavi skupaj z izrekom besede 'ste'. Oblika ob koncu sekvence udarca omogoča eno od možnih fizičnih predstavitev NCI, ki cilja na sogovornika, tj. na osebo, s katero je v neposredni komunikaciji. Zadnja, tretja gesta, pa je prej utrip kot udarec. Kot mašilo je bil izbran, da sledi prozodični strukturi govorjene vsebine.

Za vsako podštudijo smo odzive ovrednotili z eksperimentom dojemanja. Sodelujoče smo prosili, naj ocenijo odvisne spremenljivke (glej Tabelo 1), tako da opišejo kakovost predstavitve, in sicer na 5-stopenjski lestvici Likert. Poleg kakovosti gest (npr. oblika, dinamika, tekočnost, sinhronizacija) pa so prav tako opazovali, kako razumljiva je bila predstavljena vsebina. Ob upoštevanju obeh primerov (besedilo in govor ter ECA z gestami) so sodelujoči opredelili splošno dojemanje delovanja in podobnosti človeku, izraženo z zadnjo, sedmo, odvisno spremenljivko, ki smo jo prav tako ovrednotili na 5-stopenjski lestvici Likert. Omenjenih sedem meril, ki smo jih ocenjevali, je navedenih v Tabeli 1.

Tabela 1: Lestvica Likert z vrednostmi za kontekstno odvisno ovrednotenje večmodalnega izhoda

Odvisna spremenljivka	**Vprašanje**	**Lestvica**
Ujemanje vsebine (C1)	*Ali geste pravilno tolmačijo jezikovne informacije?*	1 – ne
		5 – zelo verjetno tolmačenje
Sinhronizacija oblike (C2)	*Se vam zdijo oblike v različnih govornih segmentih primerne?*	1 – ne
		5 – zelo verjetna korelacija
Tekočnost (C3)	*Je bilo gibanje tekoče?*	1 – ne
		5 – zelo tekoče
Dinamika (C4)	*Ali je bila hitrost predstavljene vsebine primerna?*	1 – prepočasi
		5 – prehitro
Zgoščenost (C5)	*Je bilo dovolj gibanja?*	1 – premalo
		5 – preveč
Razumevanje (C6)	*Kako razumljiva je bila predstavljena vsebina?*	1 – nerazumljivo
		5 – jasno razumljivo
Živahnost (C7)	*Kako bi v splošnem ocenili izkušnjo? Se vam zdi obnašanje bolj naravno in živahno glede na običajne vmesnike (brez govora in brez ECA)?*	1 – nenaravno
		5 – blizu človeku podobnemu

Merilo ujemanja vsebine (C1) označuje, ali neverbalno obnašanje in povezane geste predstavljajo ujemajoč govorni segment, medtem ko sinhronizacija oblike (C2) ugotavlja, ali je bila izbrana ustrezna vizualna predstavitev za dan segment. Merilo tekočnosti (C3) kaže na stopnjo tekočnosti sinhroniziranih gest, gibov in prehodov. Merilo dinamike (C4) smo uporabili za ocenjevanje hitrosti gest/izgovarjave. Merilo zgoščenost (C5) kaže na razporeditev ustvarjenih gest (ali je vključenih preveč ali premalo neverbalnih elementov). Merilo razumevanja (C6) smo uporabili za preverjanje, ali je bila sintetizirana vsebina (govor ali govor in geste hkrati) jasno

razumljiva in ali so posamezni segmenti bili sintetizirani/proizvedeni tako, da so slabše razumljivi. Zadnje merilo živahnosti (C7) pa smo uporabili za preverjanje, kako, če sploh, neverbalno obnašanje prispeva k dojemanju naravnosti ECA EVA.

8 Zaključek

Naravna komunikacija zajema veliko variacij obnašanja, ki se povezujejo na dinamičen in zelo nepredvidljiv način. Vključuje tudi različne družbene in medosebne signale, ki obarvajo končni rezultat. Večmodalnost v interakciji ni le nek dodatek ali stil predstavitve informacije. Večmodalnost seže močno čez semantiko in celo čez semiotične artefakte. Močno prispeva k predstavitvi informacij kot tudi medosebni in besedilni funkciji komunikacije. V tem poglavju smo orisali pristop k samodejni sintezi bolj naravnih, človeku podobnih, odzivov, ki so ustvarjeni na podlagi pogovornega modela EVA. Predstavljen model vsebuje tri povezana in ponavljajoča se ogrodja. Prvo ogrodje obsega pogovorno analizo, s katero proučujemo spontane dialoge med več osebami, da bi ustvarili različne tipe pogovornih virov (od pravil in smernic do kompleksnih večdimenzijskih značilk). Drugo ogrodje nato vključuje vseobsegajoč algoritem za sintezo afektivnega neverbalnega obnašanja, ki temelji na naključnem in neoznačenem besedilu. V nasprotju s povezanimi raziskavami pa predlagan algoritem dopušča, da je pogovorno obnašanje poganja hkrati prozodija in besedilo ter da je oblikovano z različnimi dimenzijami situacijskih, med- in znotrajosebnih kontekstov. Predvideno obnašanje, ki je dobro sinhronizirano z jezikovno ustreznico, pa moramo predstaviti uporabniku na najbolj učinkovit način. Zato tretje ogrodje predlaganega modela vključuje realizator neverbalnega obnašanja. V našem primeru smo se odločili, da prednosti sodobnega orodja za 3D-modeliranje in igralnih pogonov združimo z najsodobnejšimi koncepti realizacije obnašanja, da vzpostavimo učinkovito in visoko odzivno ogrodje, s katerimi bi ustvarjene neverbalne izraze lahko predstavili uporabnikom z realističnimi in človeku podobnimi pogovornimi agenti s telesom. Moderni realizatorji obnašanja namreč lahko podpirajo različne parametre verodostojnosti pogovornega obnašanja, kot je raznolikost in večmodalnost načrtovanja, situacijsko zavedanje, sinteza jezikovne vsebine, sinhronizacija itd. Animacijski ogrodji, kot sta Unity in Panda 3D, pa predstavljajta močno orodje za hitro in visokokakovostno zasnovo in izvedbo virtualnih, človeku podobnih, entitet. Če sklenemo, zmožnost izražanja informacij vizualno in čustveno je pri človeški komunikaciji ključnega pomena. Posledično lahko z opredelitvijo osebnosti in

čustvenega stanja ECA tak agent postane aktiven udeleženec v pogovoru. Toda če želimo, da deluje še bolj naravno, mora biti agent zmožen tekočega in skoraj nemudnega odzivanja na situacijske sprožilce, hkrati pa mora zajemati sinhronizirane jezikovne in neverbalne kanale. Predstavljen model zato predstavlja pomemben korak na poti do bolj naravnih in človeku podobnih odzivov, ki jih ustvarja stroj.

Zahvala

Raziskavo je delno financirala Javna agencija za raziskovalno dejavnost Republike Slovenije v okviru projekta HUMANIPA – Fuzija verbalnih in neverbalnih signalov za naslednjo generacijo inteligentnih komunikacijskih vmesnikov (J2-1737).

Viri in literatura

Allwood, J. (2014). »A framework for studying human multimodal communication«. *Coverbal Synchrony in Human-Machine Interaction, ed. 1*. Boca Raton: CRC Press, str. 17–39.

Allwood, J., Ahlsén, E., Lund, J., in Sundqvist, J. (2005). »Multimodality in own communication management«. V *Proceedings from the Second Nordic Conference on Multimodal Communication*. Göteborg: Göteborg University, str. 1–20.

Birdwhistell, R. L. (2010). *Kinesics and context: Essays on body motion communication*. Pennsylvania: University of Pennsylvania Press.

Carroll, R., Peikola, M., Salmi, H., Varila, M. L., Skaffari, J., in Hiltunen, R. (2013). »Pragmatics on the page: Visual text in late medieval English books«. *European Journal of English Studies*, 17(1), str. 54–71.

Cassell, J., Bickmore, T., Campbell, L., Vilhjalmsson, H., in Yan, H. (2001). »More than just a pretty face: conversational protocols and the affordances of embodiment«. *Knowledge-based systems*, 14(1-2), str. 55–64.

Chui, K., Lee, C. Y., Yeh, K., in Chao, P. C. (2018). »Semantic processing of self-adaptors, emblems, and iconic gestures: An ERP study«. *Journal of Neurolinguistics*, 47, str. 105–122.

Church, R. B., in Goldin-Meadow, S. (2017). »So how does gesture function in speaking, communication, and thinking?«. V »Why Gesture?: How the hands function in speaking, thinking and communicating«, *Gesture Studies 7*, str. 397–412.

Ciechanowski, L., Przegalinska, A., Magnuski, M., in Gloor, P. (2018). »In the shades of the uncanny valley: An experimental study of human–chatbot interaction«. *Future Generation Computer Systems*.

Cooperrider, K. (2017). »Foreground gesture, background gesture«. *Gesture*, 16(2), str. 176–202.

Ekman, P in Friesen, W (1971). »Constants across cultures in the face and emotion«. *Journal of Personality and Social Psychology*, 17(2), str. 124–9.

Esposito, A., Esposito, A. M., in Vogel, C. (2015). »Needs and challenges in human computer interaction for processing social emotional information«. *Pattern Recognition Letters*, 66, str. 41–51.

Guaitella, I., Santi, S., Lagrue, B., in Cavé, C. (2009). »Are eyebrow movements linked to voice variations and turn-taking in dialogue? An experimental investigation«. *Language and speech*, 52(2-3), str. 207–222.

Hoek, J., Zufferey, S., Evers-Vermeul, J., in Sanders, T. J. (2017). »Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study«. *Journal of pragmatics*, 121, str. 113–131.

Kita, S., Van Gijn, I., in Van der Hulst, H. (1997). »Movement phases in signs and co-speech gestures, and their transcription by human coders«. In *International Gesture Workshop*. Springer, Berlin, Heidelberg, str. 23–35.

Kopp S, Bergmann K. (2017). »Using cognitive models to understand multimodal processes: The case for speech and gesture production«. V *The Handbook of Multimodal-Multisensor Interfaces*. New York: Association for Computing Machinery and Morgan in Claypool, str. 239–276.

Kramer, L. L., Ter Stal, S., Mulder, B. C., de Vet, E., in van Velsen, L. (2020). »Developing Embodied Conversational Agents for Coaching People in a Healthy Lifestyle: Scoping Review«. *Journal of medical Internet research*, 22(2), e14058.

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., ... in Coiera, E. (2018). »Conversational agents in healthcare: a systematic review«. *Journal of the American Medical Informatics Association*, 25(9), str. 1248–1258.

Lopez-Ozieblo, R. (2018). »Can gestures help clarify the meaning of the Spanish marker 'se'?«. *Lingua*, 208, str. 1–18.

Luger E, Sellen A. (2016). »Like having a really bad PA: The gulf between user expectation and experience of conversational agents«. V *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, str. 5286–5297.

Maricchiolo, F., Gnisci, A., in Bonaiuto, M. (2012). »Coding hand gestures: A reliable taxonomy and a multi-media support«. V *Cognitive behavioural systems*. Berlin, Heidelberg: Springer, str. 405–416).

McKeown, G., Sneddon, I., in Curran, W. (2015). »Gender differences in the perceptions of genuine and simulated laughter and amused facial expressions«. *Emotion Review*, 7(1), str. 30–38.

McNeill, D. (2008). *Gesture and thought*. Chicago: University of Chicago Press.

McNeill, D. (2016). *Why We Gesture. 'The Surprising Role of Hand Movements in Communication'*. Cambridge: Cambridge University Press.

McNeill, D., Levy, E., in Duncan, S. D. (2015). »Gesture in Discourse«. V D. Tannen, H. E. Hamilton, D. Schiffrin (urd.), *The Handbook of Discourse Analysis 2*. Oxford: Wiley-Blackwell, str. 262–289.

McTear, M., Callejas, Z., in Griol, D. (2016). *The Conversational Interface: Talking to Smart Devices*. Berlin: Springer International Publishing.

Melinger, A., in Levelt, W. J. (2004). »Gesture and the communicative intention of the speaker«. *Gesture*, 4(2), str. 119–141.

Mlakar I, Kačič Z, Borko M, Rojc M. (2017). »A novel unity-based realizer for the realization of conversational behavior on embodied conversational agents«. *International Journal of Computers*, 2, str. 205–213.

Mlakar I, Kačič Z, Rojc M. (2014). »Describing and animating complex communicative verbal and nonverbal behavior using Eva-framework«. *Applied Artificial Intelligence*, 28(5), str. 470–503.

Mlakar, I., Kačič, Z., Borko, M.., in Rojc, M. (2018). »A novel realizer of conversational behavior for affective and personalized human machine interaction-EVA U-Realizer«. *WSEAS Trans. Environ. Dev*, 14, str. 87–101.

Mlakar, I., Verdonik, D., Majhenič, S., in Rojc, M. (2019). »Towards Pragmatic Understanding of Conversational Intent: A Multimodal Annotation Approach to Multiparty Informal Interaction–The EVA Corpus«. V *International Conference on Statistical Language and Speech Processing*. Cham.: Springer, str. 19–30.

Navarro-Cerdan, J. R., Llobet, R., Arlandis, J., in Perez-Cortes, J. C. (2016). »Composition of Constraint, Hypothesis and Error Models to improve interaction in Human–Machine Interfaces«. *Information Fusion*, 29, str. 1–13.

Ochs, M., Pelachaud, C., in Mckeown, G. (2017). »A User Perception--Based Approach to Create Smiling Embodied Conversational Agents«. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(1), str. 1–33.

Opel, D. S., in Rhodes, J. (2018). »Beyond Student as User: Rhetoric, Multimodality, and User-Centered Design«. *Computers and Composition*.

Peirce, C. S. (1965). *Collected papers of Charles Sanders Peirce (Vol. 5)*. Cambridge: Harvard University Press.

Philip, P., Dupuy, L., Auriacombe, M., Serre, F., de Sevin, E., Sauteraud, A., in Micoulaud-Franchi, J. A. (2020). »Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients«. *NPJ digital medicine*, 3(1), str. 1–7.

Poria, S., Cambria, E., Bajpai, R., in Hussain, A. (2017). »A review of affective computing: From unimodal analysis to multimodal fusion«. *Information Fusion*, 37, str. 98–125.

Provoost, S., Lau, H. M., Ruwaard, J., in Riper, H. (2017). »Embodied conversational agents in clinical psychology: a scoping review«. *Journal of medical Internet research*, 19(5), e151.

Queirós, A., in da Rocha, N. P. (2018). »Ambient Assisted Living: Systematic Review«. *Usability, Accessibility and Ambient Assisted Living*, str. 13–47.

Queiroz, J., in Aguiar, D. (2015). »CS Peirce and Intersemiotic Translation«. V *International Handbook of Semiotics*. Dordrecht: Springer, str. 201–215.

Rojc M, Kačič Z. (2011). »Gradient-descent based unit-selection optimization algorithm used for corpus-based text-to-speech synthesis«. *Applied Artificial Intelligence*, 25(7), str. 635–668

Rojc, M., Mlakar, I., in Kačič, Z. (2017). »The TTS-driven affective embodied conversational agent EVA, based on a novel conversational-behavior generation algorithm«. *Engineering Applications of Artificial Intelligence*, 57, str. 80–104.

ter Stal, S., Kramer, L. L., Tabak, M., op den Akker, H., in Hermens, H. (2020). »Design features of embodied conversational agents in eHealth: a literature review«. *International Journal of Human-Computer Studies*, 138, 102409.

Trujillo, J. P., Simanova, I., Bekkering, H., in Özyürek, A. (2018). »Communicative intent modulates production and comprehension of actions and gestures: A Kinect study«. *Cognition*, 180, str. 38–51.