Akcioni model govorne komunikacije čovek-mašina

Vlado D. Delić, Milan S. Sečujski, Nikša M. Jakovljević, *Members, IEEE*

***Sadržaj* — Verbalna interakcija čovek-mašina posmatra se kao komunikacija u određenom fizičkom i psiho-socijalnom okruženju i analizira se pomoću komunikacionih modela. U ovom radu analizirana je u svetlu linearnog (akcionog) modela komunikacija. Predstavljena je i jedna modifikacija modela koja ga približava interpersonalnim komunikacijama. Šenonov matematički model komunikacija iskorišćen je za modelovanje generisanja i prenosa govora u jednom smeru, kao i izvora varijabilnosti govornog signala koji ograničavaju tačnost i robusnost automatskog prepoznavanja govora.**

***Ključne reči* — ASR, govorna komunikacija, govorne tehnologije, interakcija čovek-mašina, modelovanje, TTS.**

I. Uvod

GOVORNA komunikacija čovek-mašina bazirana je u osnovi na automatskom prepoznavanju govora - ASR. Poslednjih godina ova tema postaje sve aktuelnija [1]-[3]. Analiziraju se integralno svi aspekti govorne komunikacije čoveka i mašine kao kooperativnih sagovornika u multimedijalnoj i multimodalnoj konverzaciji [4]-[7].

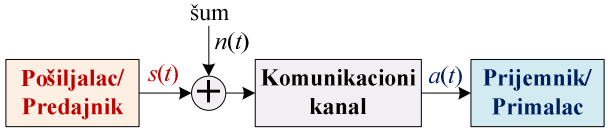
Iako čovek gotovo savršeno razume rečenice čak i ako je nivo signala 3dB (duplo) niži od nivoa buke [6], konverzacioni govor često nije dovoljno razumljiv za ASR. Da bi se uspešno realizovala verbalna interakcija čovek-mašina potrebno je dobro poznavati mogućnosti govornih tehnologija (ASR i TTS) i ukomponovati ih u okruženje. Poželjno je kreiranje pogodnih modela uz oslanjanje na iskustva iz interpersonalne usmene komunikacije [6]-[8].

Poznati modeli iz oblasti teorije komunikacija pojednostavljuju prikazivanje kompleksnih interakcija između učesnika u komunikacionom procesu. Osnovni cilj ovog rada je da kroz interpretaciju i vizuelizaciju poznatog Šenonovog modela doprinese razumevanju kompleksnih procesa govorne komunikacije čovek-mašina.

A. Šenonov i Berlov linearni model komunikacija

Klod Šenon i Voren Viver su 1949. god. opisali komunikaciju kao linearni proces. Inspirisani ondašnjom telefonskom i radio tehnologijom, razvili su model koji konzistentno objašnjava prenos informacija različitim kanalima. To je linearni model: informacije se prenose u jednom smeru, od *pošiljaoca* (izvora informacija), preko *predajnika* koji šalje poruku (signal) kroz *komunikacioni kanal*, do *prijemnika* i *primaoca* koji interpretira smisao poruke.

V. Delić, M. Sečujski, N. Jakovljević, Fakultet tehničkih nauka, Trg D. Obradovića 6, 21000 Novi Sad, Srbija (telefon: +381-21-485-2533; faks: +381-21-475-2997; e-mail: vdelic@uns.ns.ac.yu).



Sl. 1. Proces generisanja i prenosa govora.

Ovo je matematički model: elementi duž komunikacionog kanala imaju svoje prenosne karakteristike koje se u frekvencijskom domenu jednostavno množe i linearno uobličavaju spektar signala koji se prenosi. Ova linearna izobličenja i *šum* koji se u komunikacionom kanalu superponira na prenošeni signal doprinose da se manje ili više razlikuju poslata i primljena poruka, odnosno signali koji prenose ove poruke, $a(t) \neq s(t)$.

David Berlo je 1960. približio Šenonov model intrapersonalnim komunikacijama – uključujući psiho-lingvističke elemente usmene, pismene i elektronske komunikacije. Proširio je koncept izvora i odredišta sa komunikacionim iskustvom, stavovima i znanjem, socijalnim i kulturnim okvirima. Poruku je vezao za sadržaj, obradu, strukturu i kod, a kanal za ljudska čula [9].

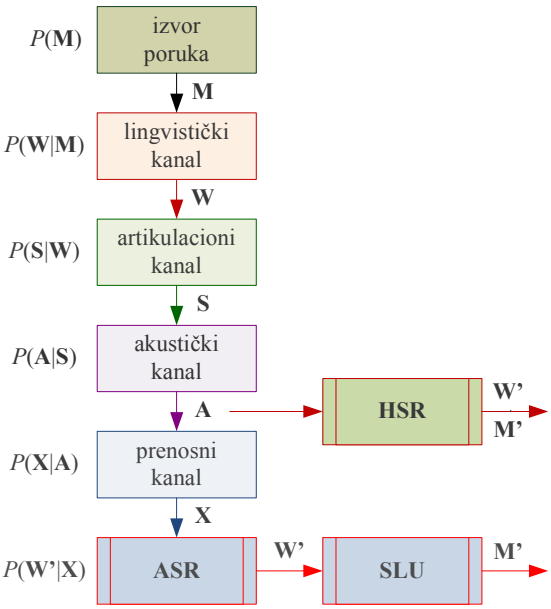
II. Varijabilnosti govornog signala

Kontinualni govor predstavlja niz izgovorenih fonema. Svaki fonem ima svoje specifične akustičke karakteristike (obeležja) koje zavise i od konteksta – koartikulacija. Kako se u govoru nižu glasovi, tako se smenjuju specifična akustička obeležja. To je izvor *varijabilnosti* (razlika) na osnovu kojih se vrši ASR. Nažalost, u govornom signalu koji dospeva na ASR, prisutan je i niz drugih varijabilnosti koje potiču od različitih govornika, ambijentalne buke, smetnji i šumova, a koje znatno otežavaju ASR. Zadatak ASR je da ignoriše neželjene izvore varijabilnosti, estimira relevantna akustička obeležja i na osnovu njih prepozna niz izgovorenih fonema (glasova).

U ovom delu rada analizirani su uzroci varijabilnosti govornog signala, kako oni na osnovu kojih se vrši automatsko prepoznaje ŠTA, KO i KAKO je rekao, tako i oni koji ograničavaju tačnost ovih automatskih procesa. Cilj ovih analiza je da se objasne neki od osnovnih modela, algoritama i parametara koji se koriste u automatskom prepoznavanju i sintezi govora (ASR i TTS).

A. Modelovanje generisanja i prenosa govora

Na Sl. 2. prikazan je proces generisanja govora i prenosa kroz zvučni ambijent do sagovornika (HSR), odnosno do mikrofona i dalje kroz prenosni kanal do ASR sistema.



Sl. 2. Proces generisanja i prenosa govora.

P(**M**) – verovatnoća poruke **M**; koje poruke će se izgovarati najviše zavisi od **aplikacije**.

P(**W**|**M**) – koje reči **W** i sekvence reči će se koristiti da izraze osmišljene poruke **M** najviše zavisi od **jezika** (*inter-language variability*), ali i u okviru jednog jezika ista poruka može se izraziti različitim rečenicama (*intra-language variability*).

P(**S**|**W**) – kako će zvučati izgovorena rečenica **W** zavisi od **govornika** (*inter-speaker variability*), pa čak ni isti govornik ne izgovara reči uvek baš na isti način (*intra-speaker variability*). U govoru **S** su pored lingvističkih (ŠTA je rečeno – **W**) prisutne i paralingvističke informacije (KO je to rekao i KAKO).

P(**A**|**S**) – artikulisani govor **S** od usta govornika do uha slušaoca i/ili mikrofona stiže direktnim putem, ali i u vidu brojnih refleksija zvučnog talasa zavisno od ambijenta – **reverberacija**. U prijemnik (mikrofon ili uho) ulazi govor (direktni i reflektovani), ali i **ambijentalna buka** (svi ostali zvuci prisutni u zvučnom polju), što dalje usložnjava zadatak ASR. Akustički signal **A** koji stiže u prijemnik zavisi kako od artikulisanog govora, tako i od akustičkog ambijenta – akustička i ambijentalna varijabilnost.

P(**X**|**A**) – Akustički signal **A** koji dospeva na ASR zavisi i od kvaliteta i položaja **pretvarača** (mikrofon, spikerfon, mobilni telefon, mikrofonski niz), kao i od **komunikacionog kanala** (npr. telefonski) kroz koji prolazi pre nego što se iz njega izdvoje sekvence obeležja **X** na osnovu kojih se najzad vrši ASR.

Prva razlika između ASR i HSR je da uho neposredno prima zvuk, a ASR posredstvom pretvarača i komunikacionog kanala sa dodatnim izobličenjima i šumom. Spoljašnje uho (školjka i kanal) pojačava govorne frekvencije, a srednje uho (kohlea) vrši spektralnu analizu u toku vremena. Mozak prati smisao, uočava izgovorene reči u kontinualnom govoru na poznatom jeziku i shvata rečenicu i govornu poruku – ovo se dešava na visokim kognitivnim nivoima koji su još uvek nedovoljno proučeni i ne mogu se lako modelovati za potrebe ASR [8].

TABELA 1: VRSTE VARIJABILNOSTI GOVORNOG SIGNALA.

Uzrok varijabilnosti	**Posledica**
razne aplikacije	različite govorne poruke
različiti jezici	različite reči i izgovor
način izražavanja	alternativni izrazi iste poruke
razni govornici	svako drugačije govori
stanje govornika	emocije, umor, bolest, starost
reverberacija prostorije	direktan zvuk i refleksije
ambijentalna buka	ostali zvuci u zvučnom polju
vrsta pretvarača	mikrofon, spikerfon, telefon
kvalitet pretvarača	usmerenost i položaj
komunikacioni kanal	telefonski, radio, VoIP
estimirana obeležja	vrsta obeležja, način estimacije

Zadatak ASR je da ignoriše sve ove varijabilnosti i da, na osnovu dobijene sekvence akustičkih obeležja **X** i predznanja o rečniku i jeziku, proceni sekvencu reči **W'**, tj. *P*(**W'**|**X**). Razlike između **W** i **W'** iskazuju se kao WER i posledica su svih pomenutih neželjenih varijabilnosti, šumova i izobličenja govornog signala. Zadatak SLU je semantička interpretacija primljene poruke (**M'**) na osnovu rezultata ASR, tj. *P*(**M'**|**W'**). Ovde se pored akustičkih i lingvističkih predznanja koriste i okviri aplikacije.

U smislu teorije informacija zadatak STT je da iz akustičkog (govornog) signala sa npr. 64 kbps izdvoji tekst **W** koji sadrži oko 1000 puta manju količinu informacija (npr. prosečno 10-12 glasova izgovorenih u sekundi, za čije kodovanje je dovoljno po 5-6 bita). Sve ostale informacije u govornom signalu mogu se smatrati redundantnim (suvišnim) sa aspekta zadatka pretvaranja govora u tekst.

B. Parametri predobrade govora za ASR

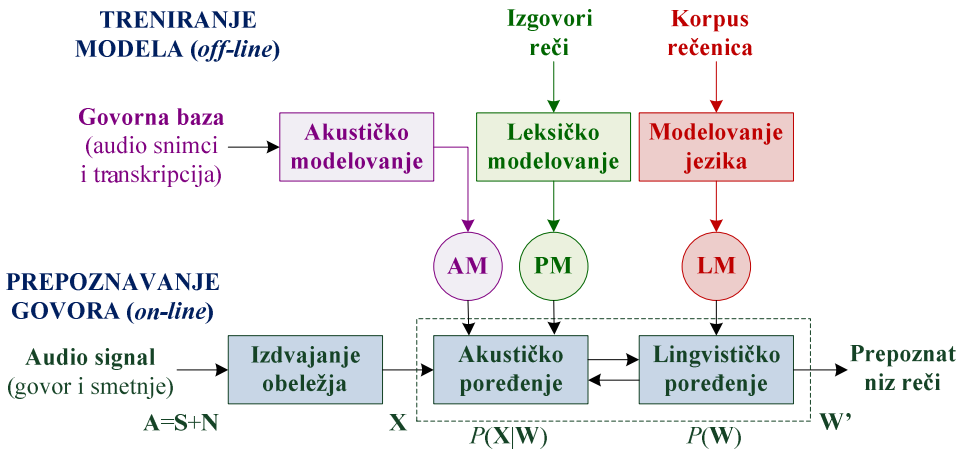
ASR algoritmi u velikoj meri baziraju se na saznanjima vezanim za percepciju govora [10].

Prvo, poznato je iz fiziološke akustike da čulo sluha u osnovi razlikuje tri karakteristike zvuka: intenzitet zvuka (glasnost), visinu tona (osnovni harmonik zvučnih segmenata) i boju zvuka (obvojnica spektra). Osnova za automatsko razlikovanje glasova jeste njihova boja, pa se zato ASR uglavnom bazira na spektralnim obeležjima govornog signala – najčešće *kepstralni koeficijenti*. Poznato je i da uho nije podjednako osetljivo na sve spektralne komponente zvuka i njihove promene duž čitavog čujnog opsega od 20 Hz do 20 kHz. Zato se koristi zakrivljenje frekvencijske ose prema tzv. melodijskoj skali, pa imamo MFCC.

Po teoriji informacija, konsonanti kao manje učestali glasovi od vokala, donose veću količinu informacije i doprinose razumljivosti govora. Međutim, njihov spektar često je izražen na višim frekvencijama koje se artikulišu sa manjim intenzitetom, čemu doprinosi NF karakter zračenja na usnama [10]. Zato su govorni signala pojačava pre izdvajanja obeležja za ASR – to je tzv. *preemfazis* – VF filtar koji izdiže više frekvencije za oko 6 dB po oktavi (toliko opada intenzitet govora na višim frekvencijama).

C. Koncept ASR na bazi HMM

Nakon svih predobrada i izdvajanja relevantnih vektora obeležja **X** počinje prepoznavanje govora koje se bazira na modularnom ili integrisanom pristupu.



Sl. 3. Modularni pristup ASR.

Integrisani pristup podrazumeva objedinjeno modelovanje svih izvora znanja na kojima se bazira ASR. Ovaj pristup ima problem što nisu podjednako dobro modelovani veoma kratki govorni segmenti (subfonemi) i veoma dugi govorni segmenti (sekvence više reči). Još uvek je suviše složen pogotovo za prepoznavanje spontanog govora.

Modularni pristup [11] kombinuje module koji zasebno modeluju različite izvore varijabilnosti: akustičke, leksičke i lingvističke, Sl. 3. Pogodnost ovog pristupa ogleda se u činjenici da stručnjaci različite struke mogu da rade na zasebnim modelima i, ako su interfejsi između modula dobro definisani, može se doći do uspešnih rezultata. Naravno, to nije lako.

Nakon *off-line* treniranja *akustičkog, leksičkog* i *lingvističkog modela* govornog signala, na ove modele se oslanja *on-line* automatsko prepoznavanje govora. Za treniranje modela potrebni su govorni i jezički resursi u vidu govornih baza i lingvističkih korpusa.

Nakon predobrade govornog signala i izdvajanja obeležja **X**, traga se za sekvencom reči **W** koja najviše odgovara opserviranoj sekvenci obeležja **X** – traži se **W** za koje je maksimalna uslovna verovatnoća $P(\mathbf{W}|\mathbf{X})$. Bajesovo pravilo prevodi zadatak u maksimizaciju obrnute uslovne verovatnoće $P(\mathbf{X}|\mathbf{W})$. Time se omogućuje automatizacija obuke i postupak ASR se svodi na postavljanje hipoteza o mogućim sekvencama reči **W**: na osnovu akustičkih i lingvističkih modela traži se ona sekvenca **W** koja maksimalno odgovara opserviranoj sekvenci obeležja **X**. To je koncept statističkog pristupa prepoznavanju govora (obično je to HMM [12]) koji je do sada pokazao najbolje rezultate.

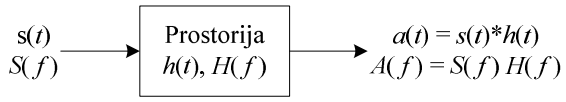
Često se zadatak ASR formalno predstavlja izrazom:

$$\mathbf{W}' = \arg\max_{\mathbf{W}} P(\mathbf{W}\,|\,\mathbf{X}) = \arg\max_{\mathbf{W}} P(\mathbf{W}) \cdot P(\mathbf{X}\,|\,\mathbf{W})$$

u kojem $P(\mathbf{X}|\mathbf{W})$ predstavlja akusički, a $P(\mathbf{W})$ lingvistički model, u skladu sa Sl. 3.

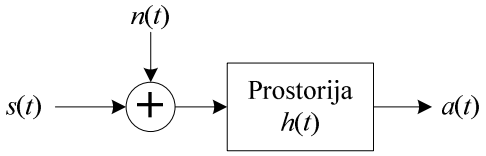
III. Linearni model za HSR i ASR

Govorni signal je obojen akustičkim ambijentom u kojem se nalaze govornik i slušalac. Svaka prostorija ima svoj *impulsni odziv* $h(t)$ iz koga se ocenjuje brzina kojom iščezava zvuk u prostoriji tj. vreme reverberacije, koje je najvažniji akustički parametar prostorije. Govorni signal $s(t)$ u linearnom modelu se konvoluira sa impulsnim odzivom prostorije, što znači da se spektar govornog signala $S(f)$ množi prenosnom karakteristikom prostorije $H(f)$ koja je Furijeova transformacija impulsnog odziva $h(t)$.



Sl. 4. Uticaj prostorije na govorni signal.

U zvučnom polju akustičkog ambijenta, pored govornog signala $s(t)$, javljaju se i drugi zvuci koji ometaju govornu komunikaciju i u linearnom modelu možemo ih prikazati kao aditivni šum $n(t)$. Neželjeni zvuci jesu *ambijentalna buka* koja može da bude raznovrsna i različitog nivoa.



Sl. 5. Uticaj ambijentalnog šuma na govorni signal.

Po Šenonovom linearnom modelu spektar audio signala $A(f)$ koji dospeva na ASR predstavlja proizvod spektra govornog signala $S(f)$ i prenosne karakteristike komunikacionog kanala $H(f)$. Tražena govorna poruka sadržana je u spektru govornog signala $S(f)$ koji za potrebe ASR treba izdvojiti iz estimirane spektralne obvojnice $X(f)$. Da bi se to pojednostavilo, treba minimizovati izobličenja signala koja su uneta u komunikacionom kanalu. Zato se spektar audio signala u toku predobrade logaritmuje (spektar se prevodi u tzv. kepstar). Pored toga što se logaritmovanjem poštuje činjenica da uho čuje logaritamski (Veber-Fehnerov zakon), proizvod spektara se prevodi u njihov zbir:

$$\log\big(S(f)\cdot H(f)\big) = \log S(f) + \log H(f)$$

Tako se stvara mogućnost da se u vreme negovorne aktivnosti estimira prenosna karakteristika kompletnog komunikacionog kanala i oduzme od zbirnog spektra – to je poznata tehnika oduzimanja prosečne vrednosti kepstrala (CMS). Preostali zadatak ASR je da se na osnovu izdvojenog spektra govornog signala $S(f)$ proceni sekvenca reči **W** koju je govornik izgovorio (ASR), odnosno poruka **M** koja se govorom prenosila (SLU). Minimizacija uticaja nestacionarne ambijentalne buke ostaje otvoreno polje

istraživanja, bez čijeg rešenja ASR sistemi nisu robustni i čim se uslovi obuke i primene ASR razlikuju, značajno se povećava greška (WER).

Čovek govori tako što artikuliše glasove pomoću vokalnog trakta koji zauzima određene konfiguracije i vrši zahvat vazdušne struje uobličavajući joj spektar – vrši se filtriranje vazdušne struje, zvučne ili šumne. Oblik spektra pojedinih glasova diktiran je parametrima filtra koji je taj spektar oblikovao. Najzad, estimacijom parametara tog filtra, ASR dobija relevantna obeležja na osnovu kojih mašina "sluša" – razlikuje pojedine glasove.

Najpoznatiji je *LPC model* vokalnog trakta koji predstavlja taj filtar kao digitalni filtar koji ima samo polove. Koeficijenti filtra menjaju se dinamikom kojom se nižu glasovi u kontinualnom govoru. Artikulacioni mehanizam je mehanički sistem koji se relativno sporo menja, tako da čovek prosečno izgovori 10-tak ili nešto više glasova u sekundi, koji traju manje od 100 ms u proseku. Zato se govorni signal ne promeni mnogo u periodima do par desetina milisekundi, pa se u tako kratkim vremenskim intervalima može smatrati približno stacionarnim. Zato se estimacija parametara filtra, tj. spektralna analiza govornog signala vrši u okviru kratkih govornih segmenata – frejmova od 30 ms koji se izdvajaju množenjem signala sa odgovarajućom prozorskom funkcijom. Da se ne bi izgubile informacije na krajevima frejmova koriste se preklopljeni prozori. Promene na prelazima između uzastopnih glasova prate se pomoću dinamičkih obeležja koja prate promene vrednosti obeležja u uzastopnim frejmovima.

Da bi se još suzio opseg uticaja neželjenih varijabilnosti u linearnom modelu govorne komunikacije, nastoji se umanjiti uticaj razlika u izgovoru koje su evidentne među različitim govornicima. Jedna tehnika za smanjenje varijabilnosti ove vrste jeste normalizacija vokalnog trakta. Već i pouzdana podela na ženske i muške govornike i kreiranje posebnih akustičkih modela za njih smanjuje WER [13].

A. Problemi koje ne pokriva linerani model

U jednosmernom matematičkom modelu uticaj ambijentalne buke bio je prosto tumačen kao aditivni šum. Time nisu rešeni određeni problemi koji se tiču govorne komunikacije, odnosno percepcije i generisanja govora.

Kad je reč o *generisanju* govora, sa povećanjem buke čovek počinje glasnije i drugačije da govori (Lombardov efekat) i to, pored same buke, otežava ASR. Isto se dešava i kada čovek stavi slušalice – čim ne čuje sebe na uobičajen način, on počinje glasnije da govori. Ovi efekti nisu obuhvaćeni linearnim modelom, ali jesu transakcionim [8] u formi iskustva sa čovekom koje je ukomponovano u dizajn sistema preko CBR modela. Poređenjem nivoa zvuka s kojim čuje sam sebe i nivoa zvuka koji čuje iz okruženja, čovek podešava intenzitet i način izgovora. Po predloženom modelu [8] tako može i mašina da se ponaša.

Što se tiče *percepcije* govora u prisustvu buke, čovek ima moć da usredsredi resurse pažnje i da se u mnoštvu zvukova koncentriše na slušanje jednog konkretnog govornog signala, a da druge potpuno ignoriše. To uspeva tako što hvata neke distinktivne karakteristike (najverovatnije pič), locira izvor (binauralno slušanje) i prati smisao govorne poruke koju sluša iz tog pravca i sa malom zadrškom shvata. Nastoje se razviti ASR sistemi koji bi po tome bili što bliži HSR, ali ovaj tzv. *coctail party* efekat nije lako modelovati pa uticaj nestacionarne buke na tačnost i robustnost ASR ostaje otvoreno polje istraživanja.

IV. ZAKLJUČAK

Govorna komunikacija u jednom smeru matematički je interpretirana pomoću linearnog modela – opisan je uticaj akustičkog ambijenta u kojem se odvija verbalna interakcija čovek-mašina, kao i drugi izvori varijabilnosti koji utiču na razumljivost govora za čoveka i mašinu.

Za tumačenje uticaja šireg psihosocijalnog konteksta na verbalnu interakciju čovek-mašina potrebni su interakcioni i transakcioni model interpersonalnih komunikacija [8].

ZAHVALNICA

Ovaj rad je podržan od strane Ministarstva za nauku i tehnološki razvoj Republike Srbije, u okviru projekta "Govorna komunikacija čovek-mašina" (TR-11001).

LITERATURA

[1] Special Issue "Spoken Language Processing", *IEEE Proceedings*, Vol. 88, No. 8, August 2000, pp. 1139-1366.

[2] Special Section "Speech Technology in Human-Machine Communication", *IEEE Signal Processing Magazine*, Vol. 22, No. 5, September 2005, pp. 16-126.

[3] Special Section "Spoken Language Technology Moves Forward", *IEEE Sig. Proc. Magazine*, Vol.25, No.3, May 2008, pp.15-97.

[4] N.O. Bernsen, H. Dybkjær, L. Dybkjær, *Designing Interactive Speech Systems*, Springer, 1998.

[5] R. A. Cole et al, *Survey of the State of the Art in Human Language Technology*, Center for Spoken Language Understanding, Oregon Graduate Institute, 1996. http://www.cse.ogi.edu/CSLU/HLTsurvey

[6] R. K. Moore, "PRESENCE: A Human-Inspired Architecture for Speech-Based Human-Machine Interaction", *IEEE Transactions on Computers*, Vol. 56, No. 9, September 2007, pp. 1176-1188.

[7] COST Action 2102: "Cross-Modal Analysis of Verbal and Nonverbal Communication", http://www.cost2102.eu

[8] V. Delić, M. Sečujski, "Transakcioni model verbalne interakcije čovek-mašina", *DOGS*, Kelebija, 2-3.10.2008, pp. 8-15.

[9] D. Mortensen, "Communication Models", Ch.2 in *Communication: The Study of Human Communication*, McGraw-Hill, 1972.

[10] S. T. Jovičić, *Govorna komunikacija, fiziologija, psihoakustika i percepcija*, Nauka, 1999.

[11] C-H. Lee, "Fundamentals and Technical Challenges in Automatic Speech Recognition", *Keynote lecture at XII SPECOM*, Moskva, 2007, pp. 25-44.

[12] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *IEEE Proceedings*, vol. 77, Feb. 1989, pp. 257-286.

[13] V. Delić, D. Pekar, R. Obradović, N. Jakovljević, D. Mišković, "A Review of AlfaNum Continuous Automatic Speech Recognition System", *in Proc. XII SPECOM*, Moskva, 2007, pp.702-707.

ABSTRACT

Speech production and transmission are analysed in this paper according to linear communicatin model. Speech signal variabilities are interpreted, as well as their impact on word error rate in automatic speech recognition. Limits of linear model in human-machine interaction are noted.

ACTION MODEL OF HUMAN-MACHINE SPEECH COMMUNICATION

Vlado D. Delić, Milan S. Sečujski, Nikša M. Jakovljević