Univerza v Ljubljani
Fakulteta za elektrotehniko

LUKS
Laboratorij za umetno zaznavanje sisteme in kibernetiko

Sedanjost in prihodnost govornih tehnologij

izr. prof. dr. Simon Dobrišek

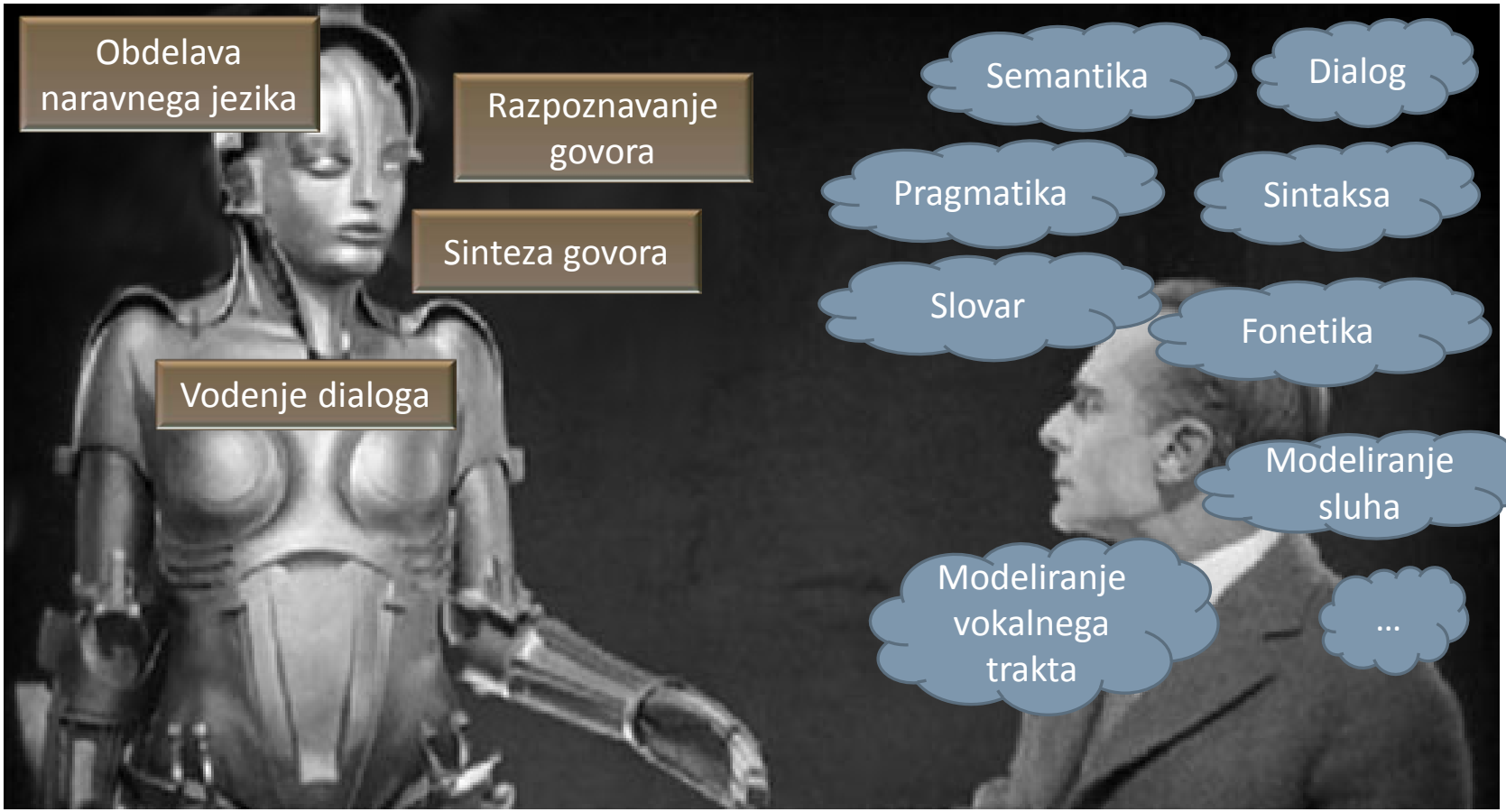
simon.dobrisek@fe.uni-lj.si

JOTA, 20. aprila 2017

Teme predavanja

- Govorne tehnologije
- Modeliranje govora
- Naše izkušnje in frustracije
- Googlova podpora govorjeni slovenščini
- Prihodnost govornih tehnologij
- Zaključni komentar

Govorne tehnologije - tradicija



Prirejeno po viru: Julia Hirschberg - Automatic Speech Recognition: An Overview

Raziskovalna področja (Interspeech)

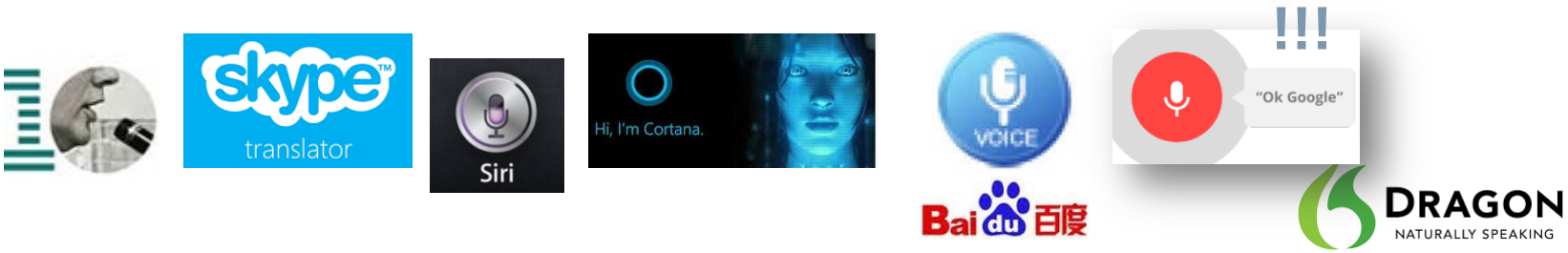
1. Zaznavanje, tvorjenje in pridobivanje govora
2. Glasoslovje in stavčna fonetika oz. prozodija
3. Parajezikovne analize govora in jezika
4. Razpoznavanje govorca in jezika
5. Obdelava govornih signalov
6. Kodiranje in izboljševanje govornih signalov

Raziskovalna področja (Interspeech)

7. Tvorjenje govora in govorjenega jezika
8. Razpoznavanje govora (akustično in jezikovno modeliranje ter aplikacije).
9. Obdelava govorjenega jezika (prevajanje, dialog, pridobivanje informacij, jezikovni viri).
10. Več-modalni govorni sistemi

Komercialni razvijalci govornih tehnologij

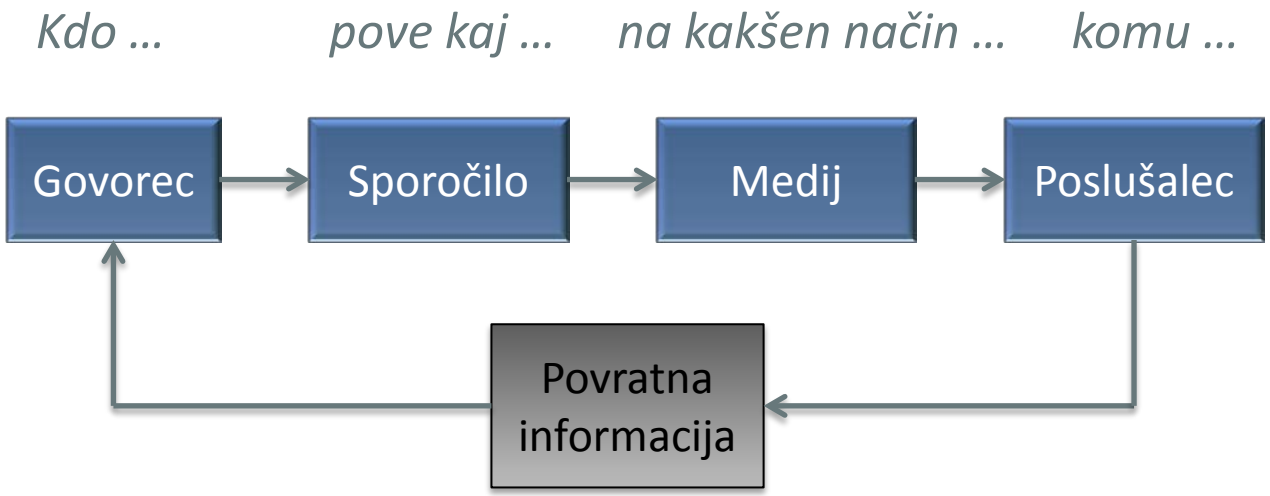
- Velika svetovna podjetja so razvila vrsto komercialnih računalniških programskih rešitev, ki vključujejo govorne tehnologije.



- Po črnogledih pričakovanjih pri teh rešitvah podpora govorjeni slovenščini še izostaja ali vsaj zaostaja.

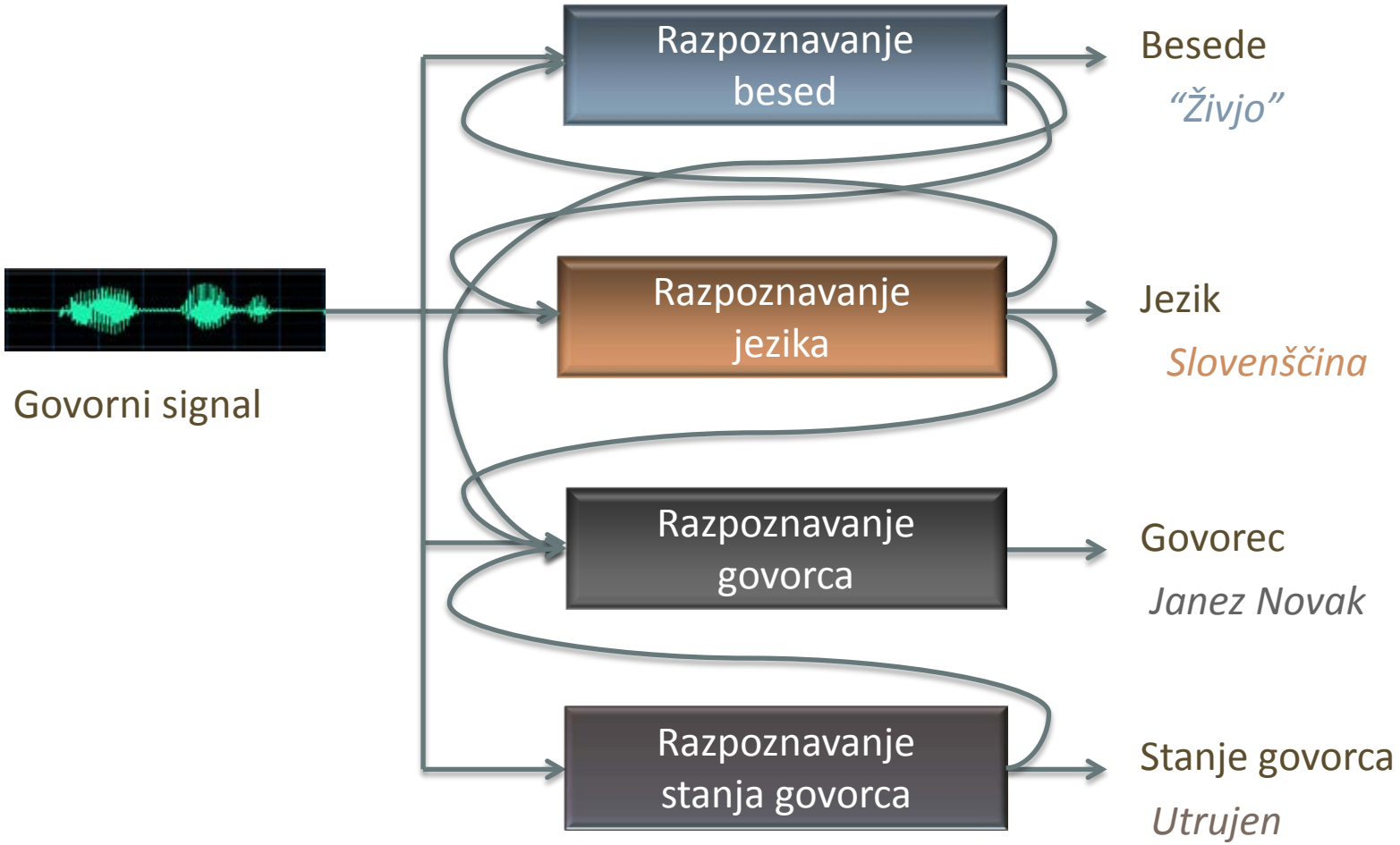
Modeliranje govora

- Govorno sporazumevanje je ena od najbolj naravnih oblik načina izmenjave informacij med ljudmi.
- Govor je predvsem besedna oblika sporazumevanja, nosi pa tudi veliko nebesednih sporočil.



- Govorno sporazumevanje je navadno dvosmerno

Luščenje informacije iz govora



Hierarhija govornih značnic/značilk

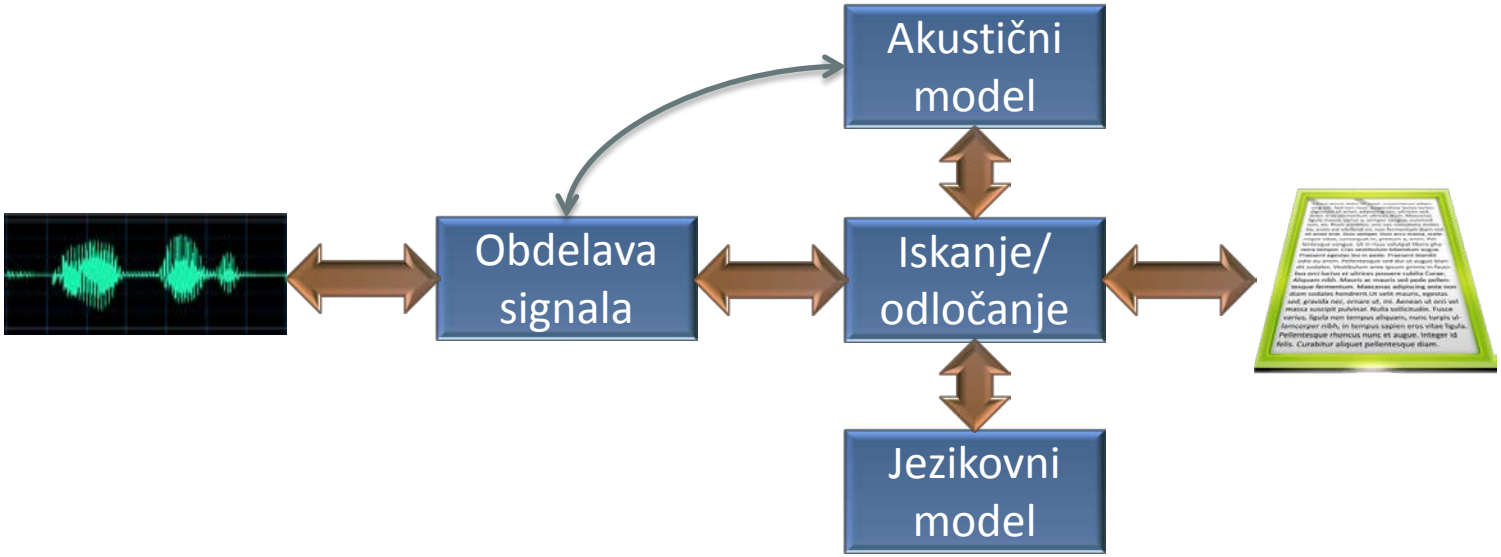
semantika, idiolekt, izgovarjava, ekscentričnost	Socialno-ekonomski status, izobrazba, kraj rojstva
prozodija, ritem, hitrost, intonacija, glasnost, melodija	osebnost, vpliv staršev
akustika, barva, globina, zadihanost, raskavost	anatomska struktura vokalnega trakta

Visokonivojske značilke (naučeno vedenje)

Nizkonivojske značilke (anatomske značilnosti)

Modeliranje govora

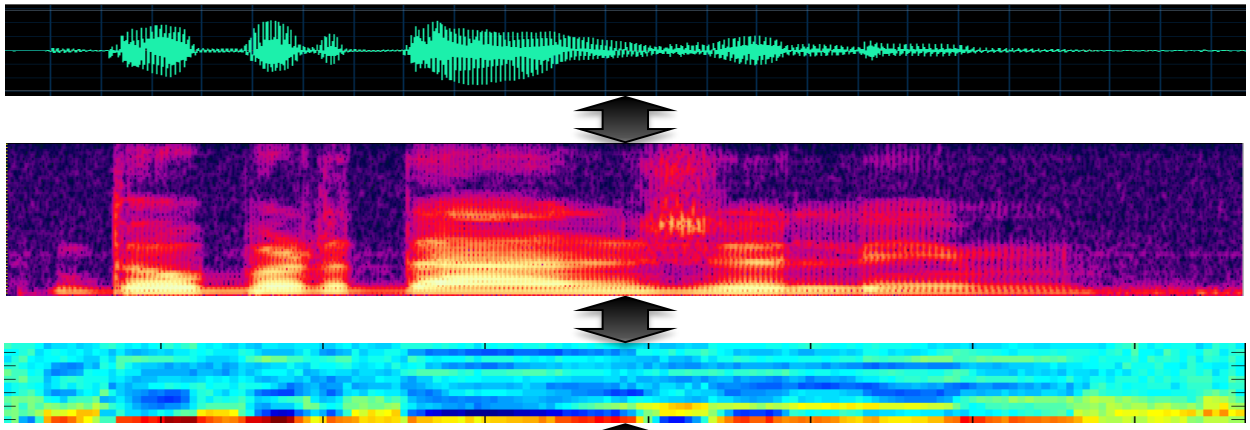
- Govorno modeliranje vključuje metode obdelave signalov, akustičnega modeliranja, jezikovnega modeliranja ter verjetnostnega iskanja in odločanja.



LUKS
Laboratorij za umetno zaznavanje sisteme in kibernetiko

Tradicionalno modeliranje govora

- Človeški govorni komunikacijski proces matematično modeliramo kot zaporedje verjetnostnih preslikav med nizi končnih stanj.



[sis],[d],[o:],[b],[@],[r],[d],[a:],[n],[Z],[E],[l],[i:],[m],[sie]

/d/,/o/,/b/,/@/,/r/,/d/,/a/,/n/,/Z/,/E/,/l/,/i/,/m/

dober,dan,želim

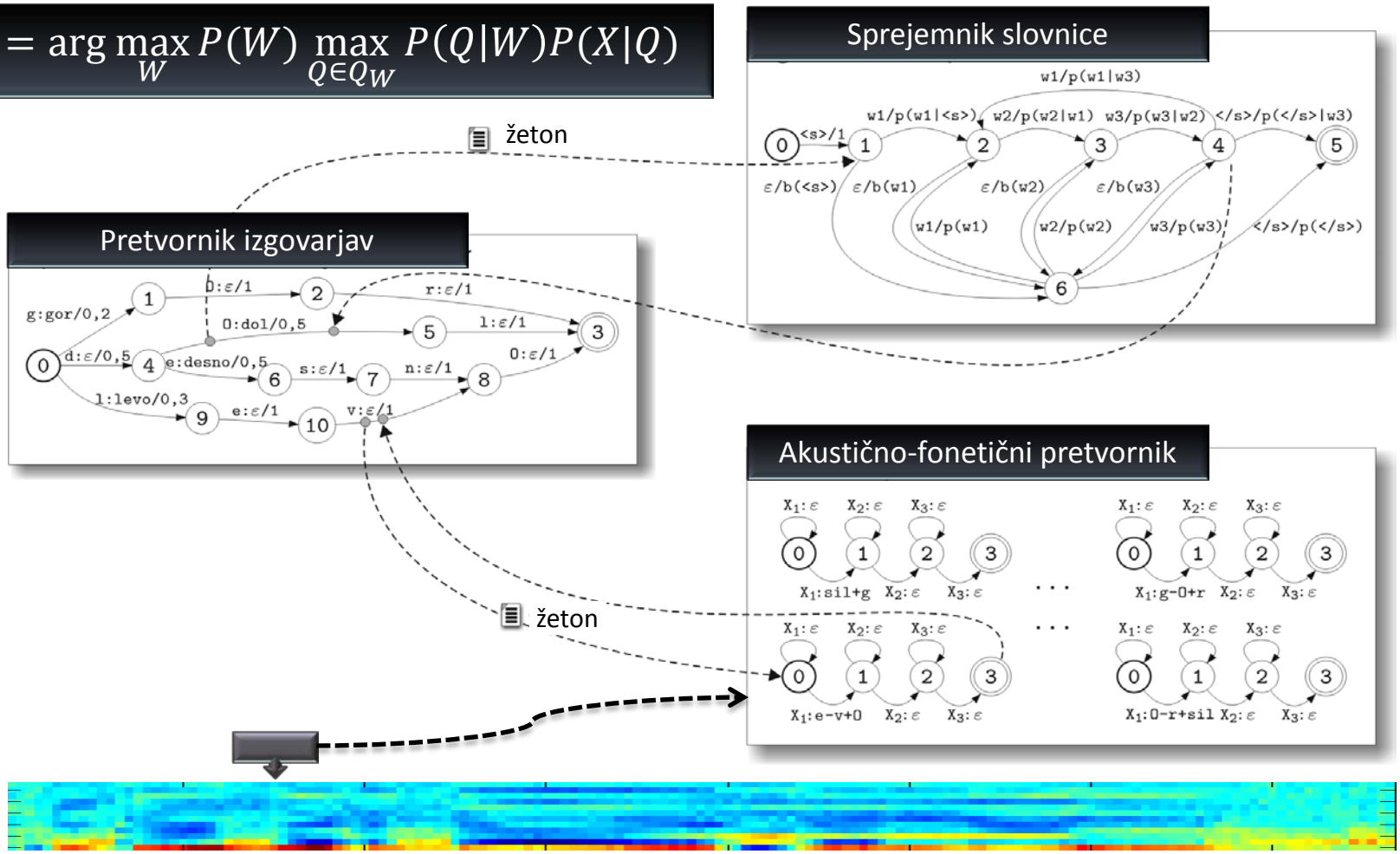
Dober dan želim !

<item>dober<tag>kakovost=pozitivno</tag></item> <item>dan<tag>čas=24h</tag></item> <item>želim<tag>…

Tradicionalno modeliranje govora

- Za modeliranje pretvorbe govora v najbolj verjeten niz besed se uporablja kaskada deterministič nih, ne-deterministič nih in verjetnostnih končnih avtomatov.

$$\hat{W} = \arg\max_{W} P(W) \max_{Q \in Q_W} P(Q|W)P(X|Q)$$



Uvajanje globokih modelov

- Uvajanje globokih akustični modelov izboljšuje modeliranje variabilnosti govora ter modeliranja kompleksnih porazdelitev akustičnih značilk.
- Uvajanje globokih jezikovnih modelov izboljšuje modeliranje sintakse, pragmatike, semantike in drugih širših jezikovnih kontekstnih odvisnosti.
- Dosežene znatne izboljšave z uporabo različnih globokih nevronskih omrežij (CNN, RNN, LSTM, nevronske „avtoceste“).
- Poskuša se tudi s t.i. „end-to-end“ modeli, brez izrazitega strukturnega ločevanja posameznih pod-modelov.

Napredek pri razpoznavalnikih govora

- Preizkus na angleški govorni zbirki „Switchboard“ (telefonski pogovori)

Sistem	Delež napačno razpoznanih besed
1995 – t.i. “Visoko zmogljiv HMM razpoznavalnik”	45%
2000 – Univeza v Cambridgeu (HTK)	19.3%
2004 – Sistem IBM	15.2%
2015 – Sistem IBM (globoki modeli)	8%
Ocena človeške zmogljivosti	4%

Naše izkušnje in frustracije

- Na Fakulteti za elektrotehniko Univerze v Ljubljani poteka razvoj govornih tehnologij že nekaj desetletij.
- V tem času smo pridobili več govornih zbirk ter razvili prve razpoznavalnike in sintetizatorje govorjene slovenščine ter prve demonstracijske govorne sisteme za pridobivanje informacij.
- V zadnjem desetletju pa smo se posvečali predvsem razvoju sistemov za razpoznavanje govorcev in njihovih psihofizičnih stanj ter sistemom za tvorjenje umetnega čustvenega govora.

Razpoznavalniki govora

- Pri razvoju lastnih razpoznavalnikov govora smo sledili razvojne trende vse do uvedbe globokih akustičnih in jezikovnih modelov, ki za učenje zahtevajo zelo velike količine podatkov.

- Z razpoznavalnikom govora, ki se prilagaja govorcu, smo na zbirki Sofes dosegli ocenjeni delež napačno razpoznanih besed 18,9%.

- Globoke modele smo zaenkrat preizkusili samo pri problemu samodejnega razpoznavanja glasov govorjene slovenščine (ocenjeni delež napačno razpoznanih glasov: 21% -> 10%).

- Sistematično smo preizkusili Googlov najnovejši govorni vmesnik, ki temelji na uporabi globokih modelov ter podpira govorjeno slovenščino.

Primer uporabe globokih modelov

```
%WER 31.4 | exp/mono/decode_dev/score_5/ctm_phn.filt.sys
%WER 34.0 | exp/mono/decode_test/score_5/ctm_phn.filt.sys
%WER 20.2 | exp/tri1/decode_dev/score_10/ctm_phn.filt.sys
%WER 23.4 | exp/tri1/decode_test/score_10/ctm_phn.filt.sys
%WER 18.7 | exp/tri2/decode_dev/score_9/ctm_phn.filt.sys
%WER 21.5 | exp/tri2/decode_test/score_8/ctm_phn.filt.sys
%WER 14.1 | exp/tri3/decode_dev/score_10/ctm_phn.filt.sys
%WER 18.4 | exp/tri3/decode_dev.si/score_10/ctm_phn.filt.sys
%WER 17.3 | exp/tri3/decode_test/score_7/ctm_phn.filt.sys
%WER 21.6 | exp/tri3/decode_test.si/score_7/ctm_phn.filt.sys
%WER 14.0 | exp/tri4_nnet/decode_dev/score_5/ctm_phn.filt.sys
%WER 17.5 | exp/tri4_nnet/decode_test/score_5/ctm_phn.filt.sys
%WER 11.8 | exp/sgmm2_4/decode_dev/score_7/ctm_phn.filt.sys
%WER 15.5 | exp/sgmm2_4/decode_test/score_6/ctm_phn.filt.sys
%WER 11.4 | exp/sgmm2_4_mmi_b0.1/decode_dev_it1/score_6/ctm_phn.filt.sys
%WER 11.2 | exp/sgmm2_4_mmi_b0.1/decode_dev_it2/score_6/ctm_phn.filt.sys
%WER 11.2 | exp/sgmm2_4_mmi_b0.1/decode_dev_it3/score_6/ctm_phn.filt.sys
%WER 11.3 | exp/sgmm2_4_mmi_b0.1/decode_dev_it4/score_6/ctm_phn.filt.sys
%WER 14.7 | exp/sgmm2_4_mmi_b0.1/decode_test_it1/score_6/ctm_phn.filt.sys
%WER 14.7 | exp/sgmm2_4_mmi_b0.1/decode_test_it2/score_7/ctm_phn.filt.sys
%WER 14.6 | exp/sgmm2_4_mmi_b0.1/decode_test_it3/score_6/ctm_phn.filt.sys
%WER 14.7 | exp/sgmm2_4_mmi_b0.1/decode_test_it4/score_8/ctm_phn.filt.sys
%WER 10.4 | exp/dnn4_pretrain-dbn_dnn/decode_dev/score_5/ctm_phn.filt.sys
%WER 13.4 | exp/dnn4_pretrain-dbn_dnn/decode_test/score_5/ctm_phn.filt.sys
%WER 10.2 | exp/dnn4_pretrain-dbn_dnn_smbr/decode_dev_it1/score_5/ctm_phn.filt.sys
%WER 10.0 | exp/dnn4_pretrain-dbn_dnn_smbr/decode_dev_it6/score_5/ctm_phn.filt.sys
%WER 13.1 | exp/dnn4_pretrain-dbn_dnn_smbr/decode_test_it1/score_5/ctm_phn.filt.sys
%WER 13.0 | exp/dnn4_pretrain-dbn_dnn_smbr/decode_test_it6/score_4/ctm_phn.filt.sys
%WER 10.9 | exp/combine_2/decode_dev_it1/score_5/ctm_phn.filt.sys
%WER 10.8 | exp/combine_2/decode_dev_it2/score_5/ctm_phn.filt.sys
%WER 10.7 | exp/combine_2/decode_dev_it3/score_5/ctm_phn.filt.sys
%WER 10.7 | exp/combine_2/decode_dev_it4/score_5/ctm_phn.filt.sys
%WER 14.3 | exp/combine_2/decode_test_it1/score_5/ctm_phn.filt.sys
%WER 14.2 | exp/combine_2/decode_test_it2/score_5/ctm_phn.filt.sys
%WER 14.1 | exp/combine_2/decode_test_it3/score_5/ctm_phn.filt.sys
%WER 14.1 | exp/combine_2/decode_test_it4/score_5/ctm_phn.filt.sys
```

Uporaba orodij Kaldi na zbirki Sofes

KALDI

35+ različnih sistemov

Tvorjenje umetnega govora

- Pri sistemih za tvorjenje umetnega govora smo v sodelovanju z Alpineonom sledili razvojne trende vse do zadnjih poskusov z uporabo globokih modelov.

- Primeri HMM sintez govora
  - HMM – Aleš Valič

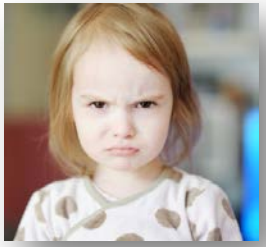
  - HMM – Tanja Cegnar

  - HMM – Janez Markošek

Razpoznavanje psihofizičnih stanj

- Doseženi ocenjeni priklic pri zaznavanju negativnih čustev otrok iz govora je bil 71,6% (govorna zbirka FAU-Aibo).





- Doseženi ocenjeni priklic pri zaznavanju alkoholiziranosti govorcev pa 70,9% (govorna zbirka VINDAT)



Razpoznavanje čustvenih stanj

- Pridobivanje govorne zbirke EmoLUKS (radijske igre).
- Upoštevanje sedmih čustvenih stanj pri označevanju govora (nevtralno, žalost, strah, jeza, gnus, presenečenje, veselje).
- Pri razpoznavanju vseh sedmih čustvenih stanj je bil dosežen ocenjeni uteženi priklic 52,8%.
- Pri zaznavanju vzbujenega stanja (vseh šest čustvenih stanj) pa je bil dosežene uteženi priklic 72,1%.
- Zbirka EmoLUKS je bila uporabljena predvsem pri razvoju sistema za tvorjenje umetnega čustvenega govora.

Umetno tvorjeni čustveni govor

Čustveno stanje	Moški glas	Ženski glas
Izvirno		
Nevtralno		
Jeza		
Veselje		
Strah		
Presenečenje		
Žalost		

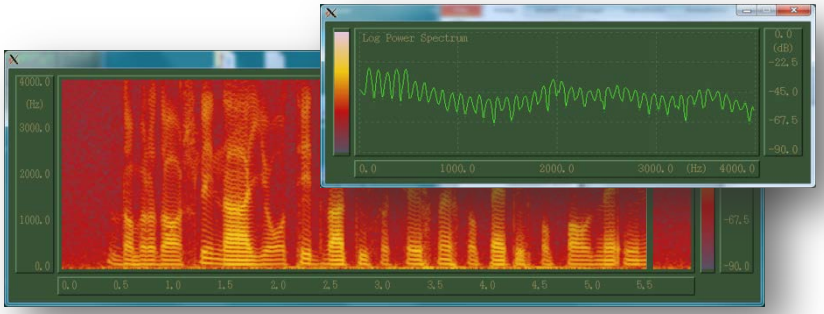
Razpoznavanje govorcev

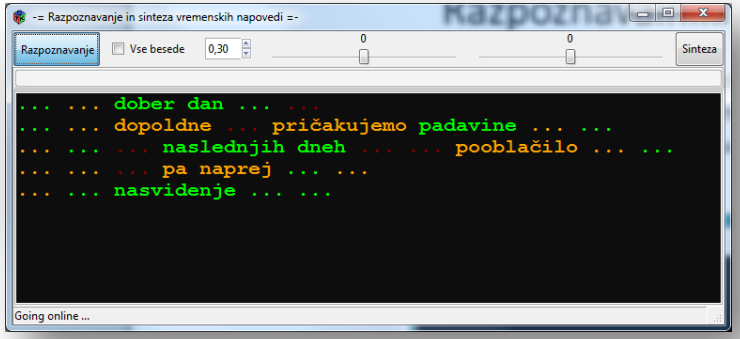
- V sodelovanju z Alpineonom je bila dosežena zmaga na tekmovanju IAPR biometrične konference ICB-2013.

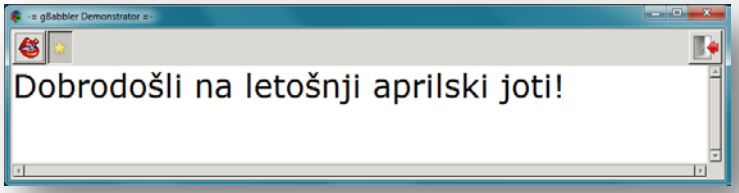
SYSTEM	FEMALE		MALE	
	DEV (EER %)	EVAL (HTER %)	DEV (EER %)	EVAL (HTER %)
Alpineon	7.982	10.678	5.040	7.076
ATVS	16.836	17.858	14.881	15.429
CPqD	14.348	15.987	11.824	10.214
CDTA	19.471	22.640	12.738	19.404
GIAPSI	11.590	12.813	9.683	8.865
EHU	17.937	19.511	11.310	10.058
IDIAP	12.011	14.269	9.960	10.032
L2F	13.484	22.140	10.599	11.129
L2F-EHU	11.005	17.266	7.889	8.191
Mines-Telecom	11.429	11.633	10.198	9.109
Phonexia	8.364	14.181	9.601	10.779
RUN	25.405	23.112	24.643	22.524

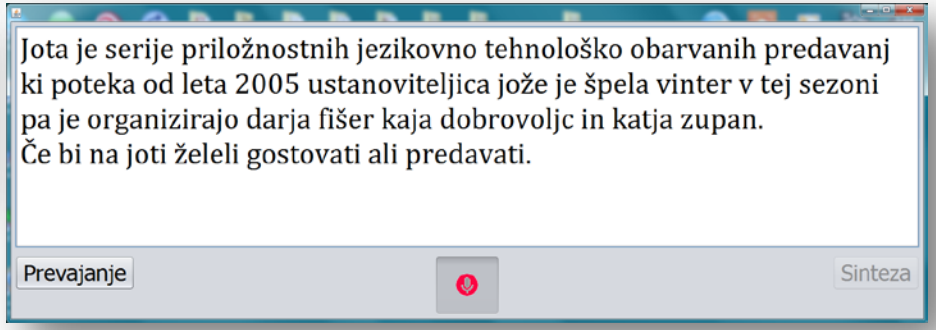
Demonstracijske aplikacije

- Razvili smo več govornih demonstracijskih aplikacij (XSFV, Simian, gBabbler, Viridian, Dialogic) ter programski vmesnik za vključitev govornega ukazovanja v različne aplikacije (Vociferator).



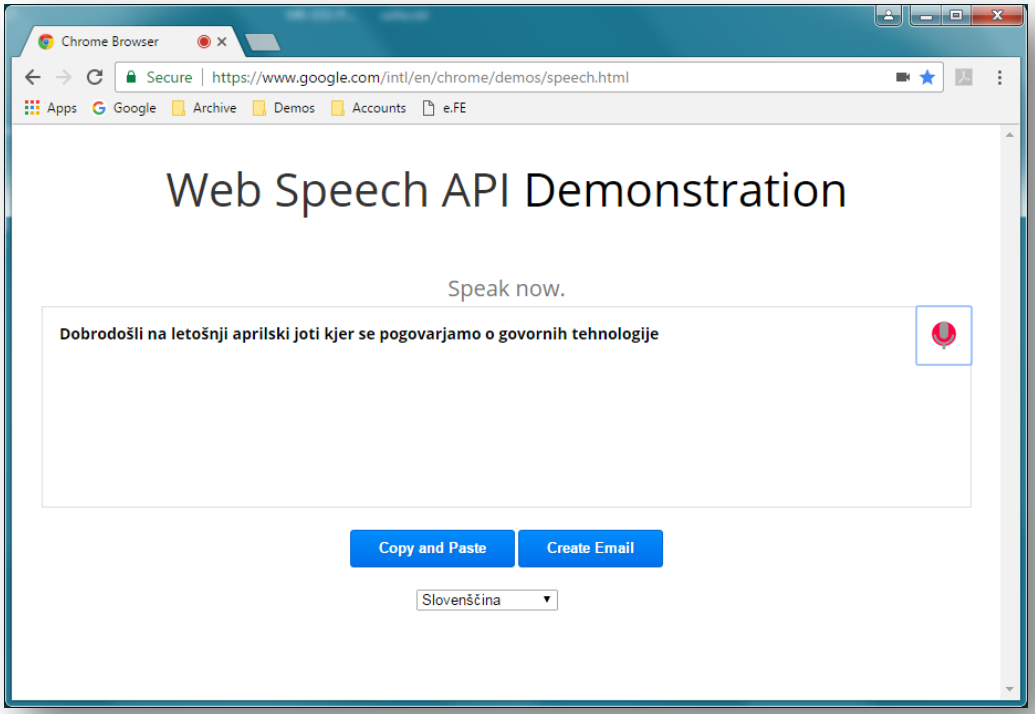






Googlova podpora govorjeni slovenščini

- Med najbolj razvitimi komercialnimi govornimi vmesniki je zaenkrat uspelo pri razpoznavalnikih govora le Googlu podpreti tudi govorjeno slovenščino.



Googlov govorni programski vmesnik

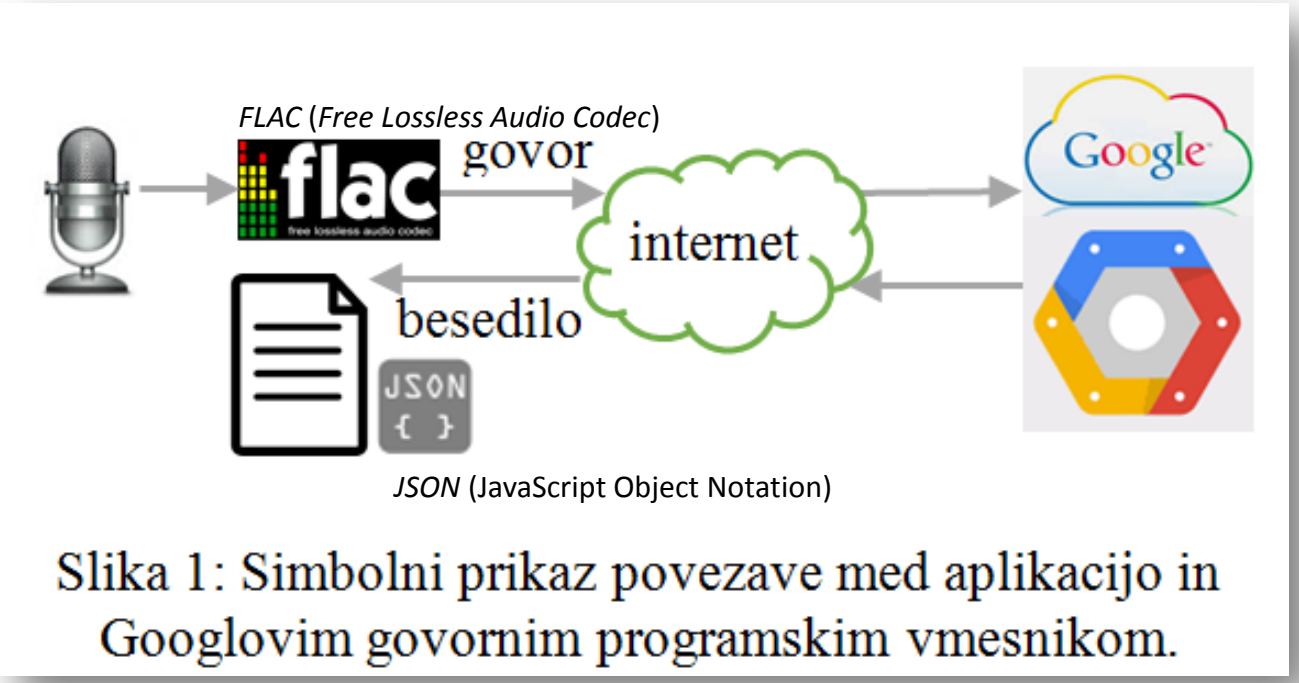
- Googlov govorni programski vmesnik je slabo dokumentiran in omejeno dosegljiv za širšo skupnost razvijalcev novih aplikacij.
- Vmesnik se najbolj enostavno uporablja z uporabo programskega jezika Javascript in programskega vmesnika Google Javascript Speech API.
- V tem primeru se aplikacije, ki uporabljajo Googlov govorni vmesnik, razvijejo kot običajne spletne aplikacije, ki se naložijo v spletni brskalnik (Chrome).

Googlov govorni programski vmesnik

- Za neposredno uporabo Googlovega govornega programskega vmesnika potrebujemo ključ, ki je registriran v Googlovem oblačnem sistemu.
- Po pridobitvi ključa, ki istoveti našo aplikacijo, lahko Googlov govorni programski vmesnik brezplačno uporabimo do 50-krat na dan in z omejitvijo na zvočne posnetke, ki so lahko dolgi do 15 sekund.
- Za intenzivnejšo uporabo govornega programskega vmesnika je potrebno plačilo po posebnem ceniku.

Googlov govorni programski vmesnik

- Google govorni programski vmesnik lahko uporabljamo na izmenjujoči enosmerni ali hkratni dvosmerni način.



Slika 1: Simbolni prikaz povezave med aplikacijo in Googlovim govornim programskim vmesnikom.

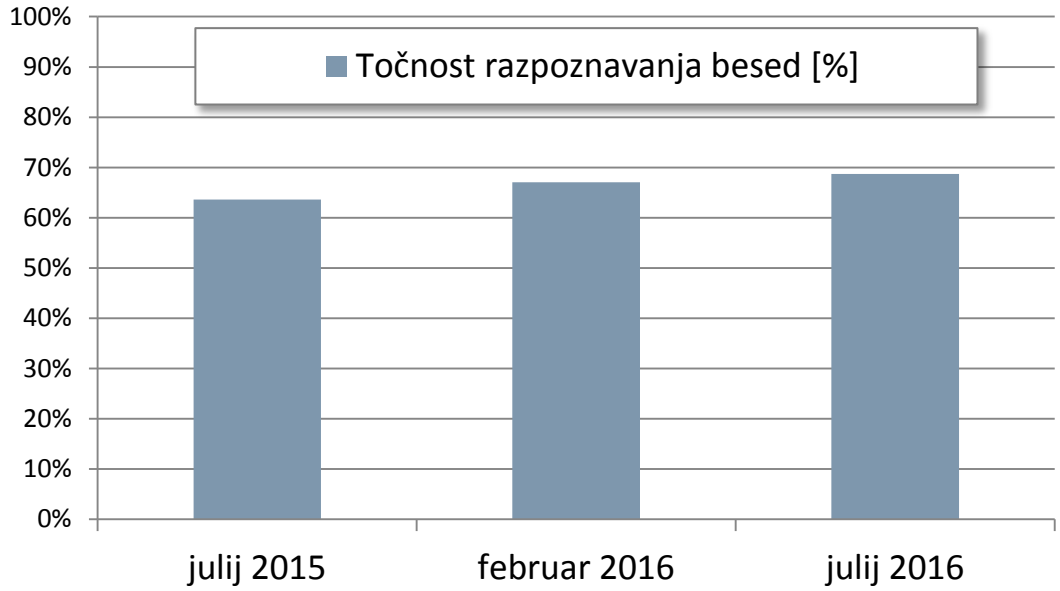
Preizkus govornega vmesnika

- Preizkus Googlovega govornega vmesnika smo izvedli predvsem z uporabo lastnih zbirk govornih posnetkov GOPOLIS in VNTV.
- Uporabili smo tudi nekaj dodatnih zvočnih posnetkov prebiranja elektronskih pisem v slovenščini (dolžine okoli 100 besed).
- Elektronska pisma so se prebirala najprej počasi in razločno, nato normalno hitro in manj razločno ter nato še spontano, manj razločno, z medmeti in s prekinitvami.

Rezultati na govorni zbirki GOPOLIS

- Skupaj 1925 testnih posnetkov povedi v skupnem trajanju dobre 1 ure in 37 minut se je poslalo Googlovemu govornemu programskemu vmesniku in pridobilo rezultate razpoznavanja.
- Preizkus smo v zadnjem letu dni izvedli trikrat, ker nas je zanimalo, če se bo rezultat zaradi morebitnega prilagajanja Googlovega razpoznavalnika govora že obdelanim govornim posnetkom v vmesnem času kaj izboljševal.

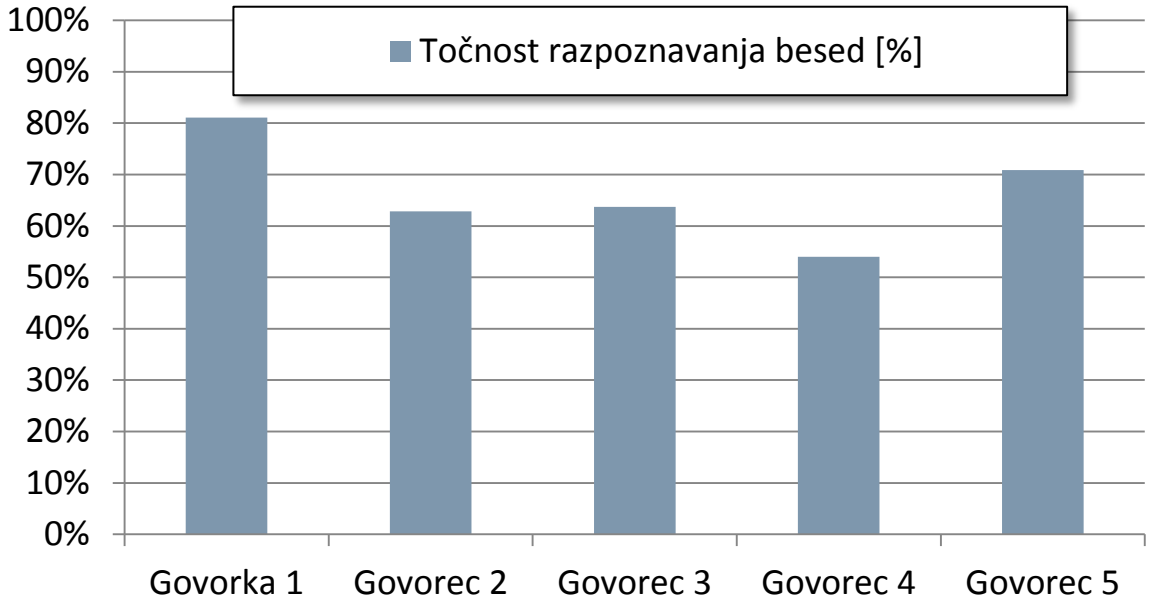
Rezultati na govorni zbirki GOPOLIS



- Zaradi večjega števila lastnih imen krajev, letališč in letalskih prevoznikov je del napak tudi posledica napak pri njihovem ortografskem zapisu, na primer *Sheremetyevo – Šeremetjevo, Zuerich – Cirih, ipd.*

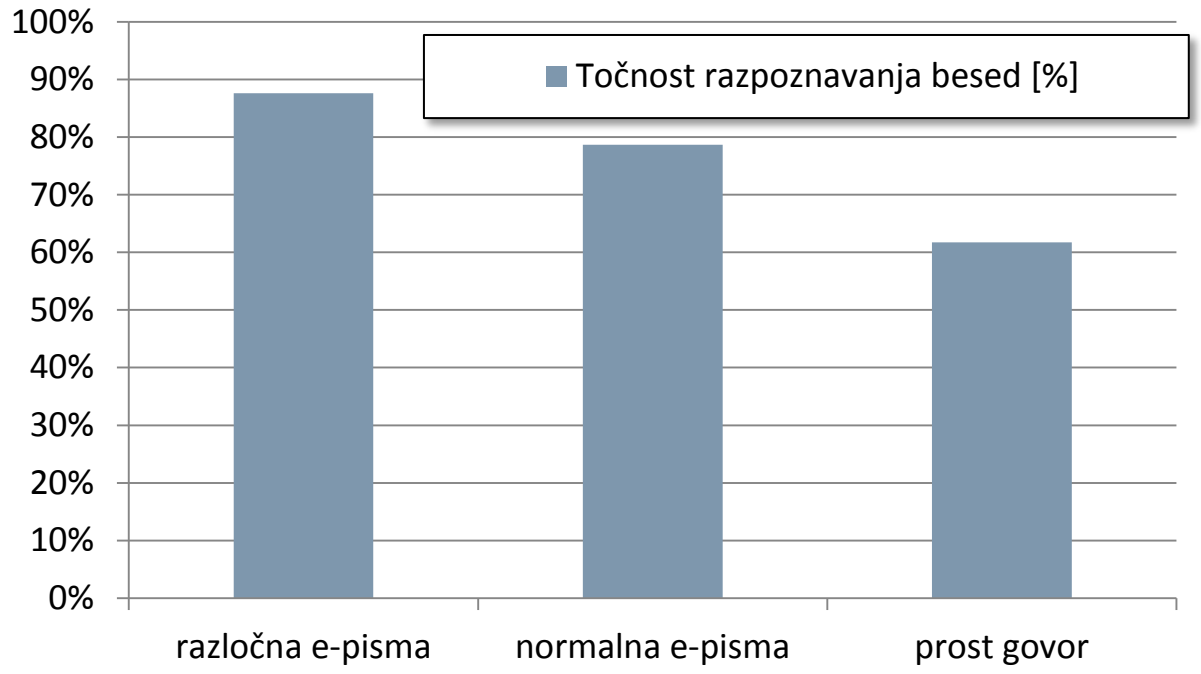
Rezultati na govorni zbirki VNTV

- Uporabili smo še 1548 televizijskih posnetkov povedi iz vremenskih napovedi petih govorcev iz govorne zbirke VNTV v skupnem trajanju 1 ure in 48 minut.



Rezultati na elektronskih pismih

- Rezultati razpoznavanja prebiranja elektronskih pisem, narekovanih na tri načine.



Nerešeni problemi govornih tehnologij

- Prilagoditev na nove okoliščine z malo razpoložljivih učnih podatkov.
- Uporaba prostorskih mikrofonov v akustičnih okoliščinah z veliko šuma, motenj in odmevov.
- Modeliranje izrazito naglašenega in narečnega govora.
- Modeliranje spontanega, netekočega in izrazito čustvenega govora.

Problem komunikacijske zasebnosti

- Govorno sporazumevanje med ljudmi načeloma omogoča večjo komunikacijsko zasebnost kot pisno sporazumevanje.
- Govorno sporazumevanje s računalniki pa to prednost izniči (možno shranjevanje posnetkov).
- Problem je tudi akustična govorna komunikacijska zasebnosti v fizični bližini govorca (motenje sodelavcev z glasnim govorom ipd).

Naše prihodnje raziskovalne ambicije

- Uporaba globokih akustičnih in jezikovnih modelov pri lastnih razpoznavalnikih in sintetizatorjih govorjene slovenščine (Kaldi, MS CNTK, CUED HTK, CMU Sphinx).
- Razvoj prostorskega avdio sistema z možnostjo osredotočanja na posamezne govorce ter odstranjevanje šuma in motenj

- Razvoj orodja za samo-učenje in samo-ocenjevanje govornih veščin ter pomoč pri učenju na pamet (samodejni prišepetovalec)

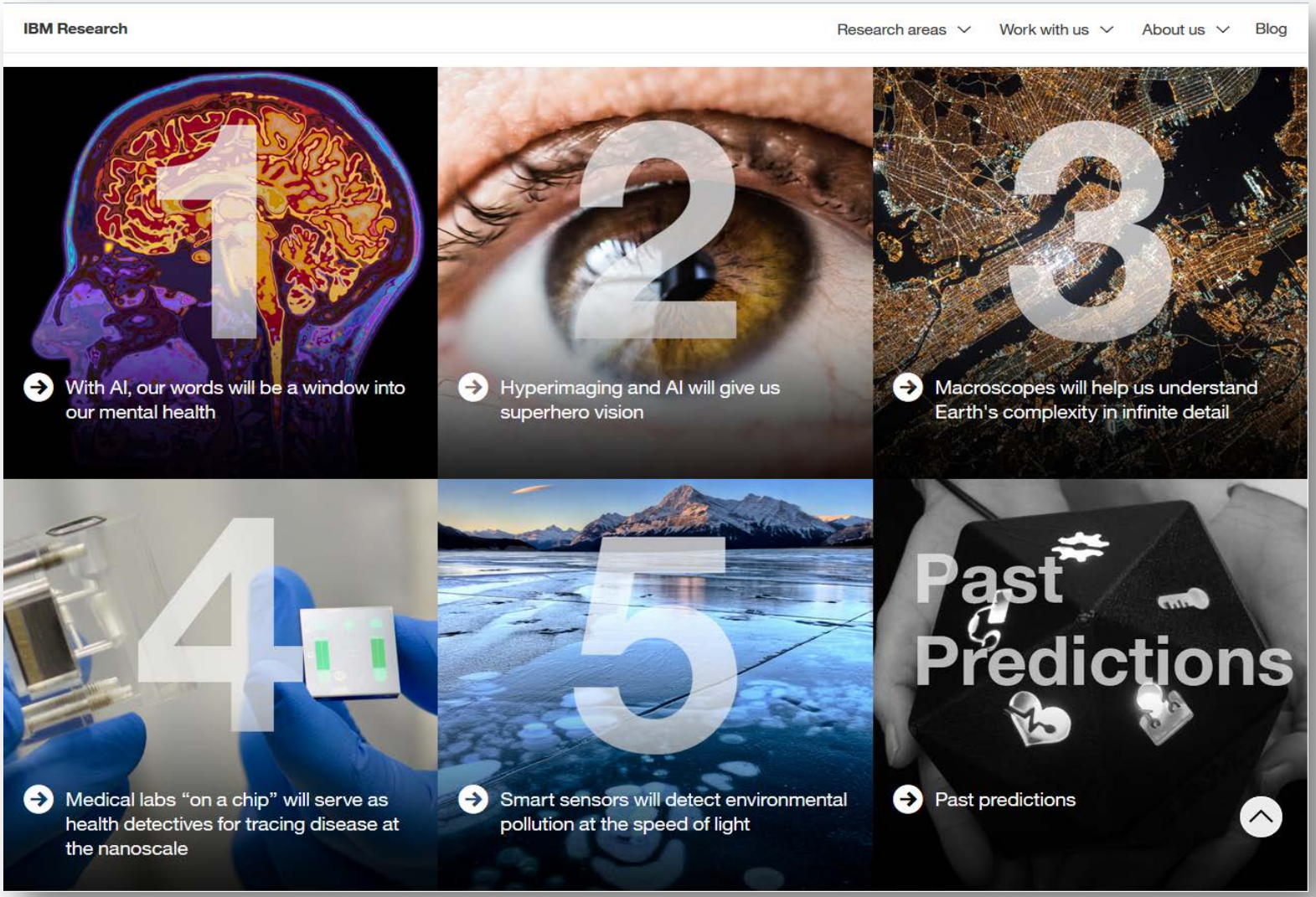
Prihodnost govornih tehnologij

- Govorna komunikacija bo še nekaj generacij človeku najbolj naraven način sporazumevanja.
- Obstaja veliko aplikacij, pri katerih je govorno sporazumevanje tudi najbolj učinkovit način sporazumevanja (ko so zasedene roke in/ali oči).
- Pri določenih govornih aplikacijah pa se odpira problem komunikacijske zasebnosti in varovanja podatkov.

Prihodnost govornih tehnologij

- Razmah umetnih inteligentnih osebnih asistentov, prevajalcev, prišepetovalcev in tolmačev.
- Poizvedovanje po informacijah med vožnjo in gibanjem.
- Vnašanje podatkov med delom in gibanjem.
- Samodejno ocenjevanje govornih zmožnosti in pomoč pri učenju.
- Govorno podprti ambientalni inteligentni sistemi.
- Razpoznavanje jezika, narečja in govorcev ter ugotavljanje njihovih psihofizičnih stanj v pametnih nadzornih sistemih.

Prihodnost govornih tehnologij



Vir: *Five innovations that will change our lives in the next five years* - http://research.ibm.com/5-in-5

Zaključni komentar

- Razvoj govornih tehnologij napreduje
- Podpora govorjeni slovenščini se izboljšuje
- Oblačne rešitve imajo svoje prednosti in slabosti
- Govorne tehnologije so vse bolj uporabne
- Nadaljnje izboljšave so vse bolj zahtevne
- Priložnosti so v ožjih raziskovalno-razvojnih nišah
- Pomembna je podpora komunikacijski zasebnosti