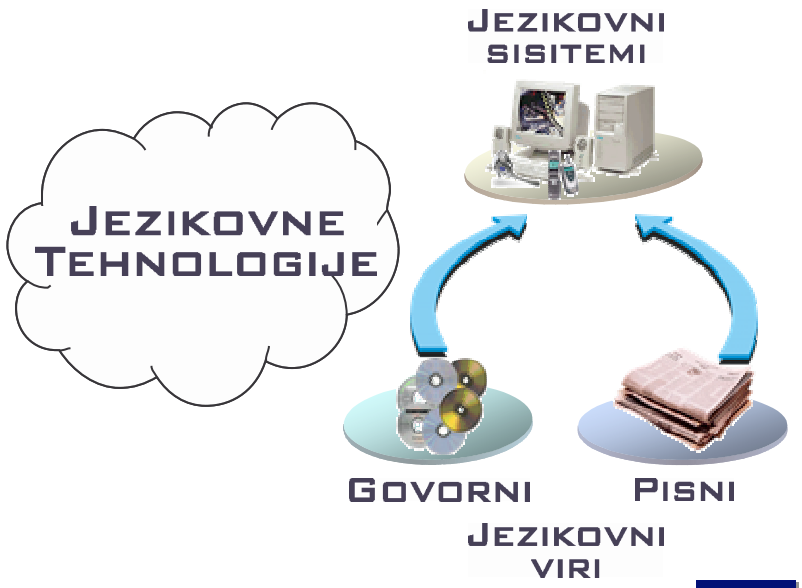
GOVORNE TEHNOLOGIJE IN JEZIKOVNI VIRI

Zdravko KAČIČ

Jezikovna raznolikost in telekomunikacijske storitve

- Na svetu je po nekaterih ocenah trenutno okrog 6000 jezikov.
- Po napovedih naj bi se jih v naslednjih 100 letih ohranila le polovica ali celo le nekaj sto.
- Med ogrožene jezike sodijo v prvi vrsti tisti, za katere obstaja le govorjena oblika, v precejšnji meri pa so ogroženi tudi jeziki majhnih narodov.
- V geografski Evropi naštejemo danes 225 različnih jezikov.

- Jezikovne tehnologije vključujejo področji govorne tehnologije in procesiranja naravnega jezika. Ni 100 % zanesljiva tehnologija.
- Govorna tehnologija vključuje sisteme
  - avtomatskega razpoznavanja govora,
  - avtomatske sinteze govora,
  - govornega dialoga,
  - strojnega simultanega prevajanja govora…
- Področje procesiranja naravnega jezika vključuje
  - črkovalnike,
  - sisteme povzemanja besedila,
  - sisteme tvorjenja besedila,
  - sisteme prevajanja besedila…



Težavnost razpoznavanja govora

- Govor je zelo spremenljiv, tudi pri istem govorcu:
  - spremenljiva krivulja višine govora (npr. emocije – presenečenje, jeza),
  - spremenljiva hitrost,
  - učinek zamašenega nosu (npr. nošenje maske).
- Različni govorci izgovarjajo besede različno:
  - različno naglaševanje vnaša velike spremembe,
  - manjše spremembe znotraj istega naglasa.
- Karakteristike glasov se spreminjajo glede na kontekst
  - spati --> stati --> stoti
- Spontan govor vsebuje za razpoznavanje govora nebistvene glasove
  - kašelj, mašila (umm, err ...), napačne začetke...

Tipologija sistemov ARG

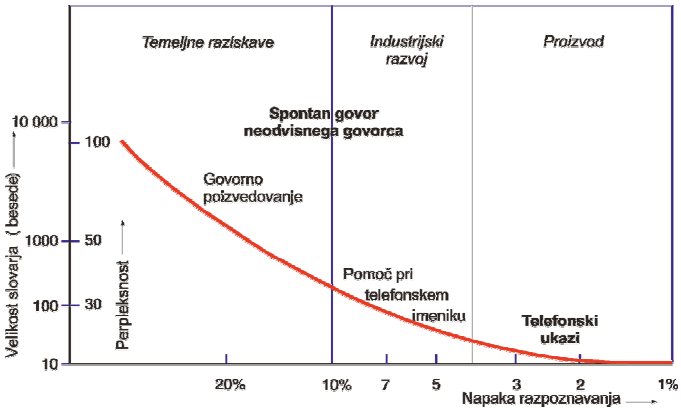
- značilnosti vezane na govorca:
  - odvisni
    - visoka uspešnost razpoznavanja
    - ni splošno uporaben
  - več govorcev
    - učna množica tvorjena z govorom potencialnih uporabnikov
  - neodvisni
    - uporabnikov govor ni v učni množici
- jezikovne omejitve:
  - razpoznavanje izoliranih besed
    - premori med besedami ≥ 200ms
  - razpoznavanje vezanih besed
    - konkatenacija izol. razpoznanih enot
    - govorec lahko uporablja tekoč govor
  - iskanje ključnih besed
  - razpoznavanje tekočega govora
    - mora upoštevati fonološke in različne koartikulacijske učinke.

- omejitve robustnosti:
  - okolje laboratorija (pisarne)
    - vsiljen mikrofon, brez okoliškega šuma
  - telefonski sistem
    - kontrola šuma znotraj nekega nivoja
  - realne razmere

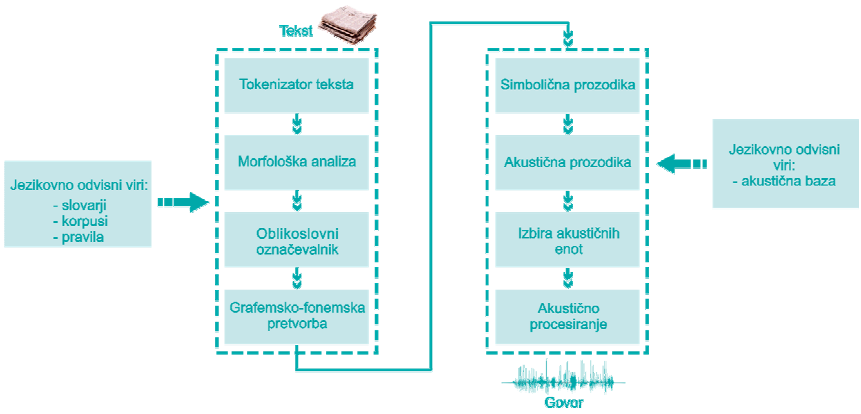
Slovar:
- majhen (1-99 besed)
- srednji (100-5000)
- velik (do 60000 in več)

- šumna okolja:
  - šum ozadja
    - v avtu, letalu
  - akustično okolje
  - komunikacijski kanal
  - govorec
    - cmokanje, dihanje...

Uspešnost razpoznavanja in uporabnost



Sistem pretvorbe besedila v govor



Jezikovna raznolikost in telekomunikacijske storitve

- Nujnost sodobnih telekomunikacij je uvajanje storitev z visoko stopnjo dodane vrednosti.
- Govorna tehnologija predstavlja eno izmed pomembnih tehnologij za dosego teh ciljev.
- Nudi možnost avtomatizacije obstoječih storitev in cenenega razvoja množice novih storitev v porajajočih se komunikacijskih omrežjih.
- Danes so najpomembnejši tržni segmenti uporabe govorne tehnologije, največkrat v obliki govorno vodenih vmesnikov, omrežni strežniki, mobilni komunikacijski terminali in potrošniške naprave.

Jezikovni viri, ki jih gradi raziskovalna skupina na FERI

Baza SpeechDat II

- Baza SpeechDat II je bila zgrajena v okviru evropskega projekta SpeechDat II Speech Databases for Creation of Voice Teleservices LE2 4001, v katerem je skupina na FERI sodelovala kot pogodbeni partner podjetja SIEMENS A.G. iz Muenchna. Za slovenski jezik je bila zgrajena baza 1000 govorcev. Bazo sestavlja 5 CD ROM-ov. Dosegljiva je preko organizacije ELRA /ELDA.

Baza Interface

- Baza Interface je govorna baza emocionalnega govora. Posneta je bila v okviru mednarodnega projekta Interface, ki je potekal v okviru 5. okvirnega programa EU. Baza je posneta v studijskem okolju. Vključuje posnetke govora šestih emocionalnih stanj: strah, jeza, veselja, gnus, žalost, presenečenje, in dveh nevtralnih stilov govora. Govor sta posnela profesionalni igralec in igralka.

Baza SNABI

- Baza izgovarjav SNABI predstavlja bazo z velikim slovarjem besed. Sestavljajo jo trije slovarji: besede, mmc in lingua. Skupaj zajemajo slovarji baze 1530 stavkov, 150 besed in abecedo. Baza vsebuje govorni signal, posnet v studiu in preko telefona, ki ga je izgovorilo 132 govorcev. Vsak izmed njih je prebral 200 ali več stavkov iz posameznih slovarjev. Vsebuje več kot 15.000 posnetkov govornega signala.

Baza PoliDat

- PoliDat je baza izgovarjav, ki je skladna s specifikacijami baze SpeechDat II. Namen projekta PoliDat je izgradnja baze izgovarjav za razvoj govorno vodenih telekomunikacijskih storitev. Baza je razdeljena v dva dela: govor, posnet preko mobilnega telefonskega omrežja, in govor, posnet preko fiksne telefonske linije. Baza bo vsebovala govor 1000 govorcev za vsak del baze (mobilni, fiksni).

Baza Broadcast News za slovenski jezik

- Bazo Broadcast News za slovenski jezik razvijamo v okviru projekta z Radiotelevizijo Slovenija. Omogočala bo razvoj sistemov razpoznavanja tekočega in pogovornega govora za slovenski jezik. Oddaje, ki bodo vključene v bazo so bile izbrane v skladu z obstoječimi priporočili za gradnjo tovrstnih baz (dolžina oddaj, zvrst, snemalno okolje …) in pokrivajo časovno obdobje od 1999 do 2003. Baza vključuje 35 ur govornega materiala, ki je obdelan in transkribiran skladno s priporočili mednarodnega konzorcija za baze Broadcast News. Vključuje besedilni korpus iNEWS iz arhiva RTV SLO.

Glasoslovni slovar Onomastica

- Onomastica je glasoslovni slovar slovenskih lastnih imen. Zgrajen je bil v okviru mednarodnega projekta EU Copernicus. Zajema popis vseh slovenskih lastnih imen (po evidenci Urada za statistiko R Slovenije iz leta 1995) z oznakami o vrsti lastnega imena in fonetični prepis.

Jezikovni viri projekta LC-STAR

- Cilj evropskega projekta 5. O.P. LC-STAR - Lexica and Corpora for Speech to Speech Translation Components je zagotoviti standardizirane jezikovne vire za razvoj tehnologije strojnega simultanega prevajanja govora (13 jezikov). Skupina na FERI kot partner projekta razvija govorne in pisne jezikovne vire za razvoj te tehnologije za slovenski jezik (glasoslovni, oblikoslovni in vzporedni slovarji) .

TURDIS – govorni korpus spontanega govora

- Razvoj govornega korpusa spontanega govora s področja turističnih informacij. Namen: razvoj pilotskih projektov sistemov strojnega simultanega prevajanja govora in analiza diskurza. Zagotoviti čim bolj realne razmere. Cilj: 200 dialogov. Trenutno stanje: 70 dialogov.

Oblikoslovni slovar SImlex

Glasoslovni slovar SIflex

Besedilni korpus Večer

Zaključek

- Uporaba govorne tehnologije omogoča razvoj telekomunikacijskih storitev z visoko stopnjo dodane vrednosti.
- Njihova širša uporaba predvidena v naslednjih nekaj letih.
- Nujnost zagotavljanja potrebnih govorjenih in pisnih jezikovnih virov.
- Zaradi nacionalnega pomena moramo nujno zagotoviti ustrezne standardizirane pisne in govorjene jezikovne vire, ki bodo zagotavljali razvoj govornih tehnologij za slovenski jezik.