

Kõnekeele ressursid ja kõnetehnoloogia andmebaasid

Einar Meister

TTÜ Küberneetika Instituut

Ülevaade ettekandest

- Eesmärgid
- Olulisus
- Senised tulemused
- Edasine töö
- Täitjad, finantseerimine

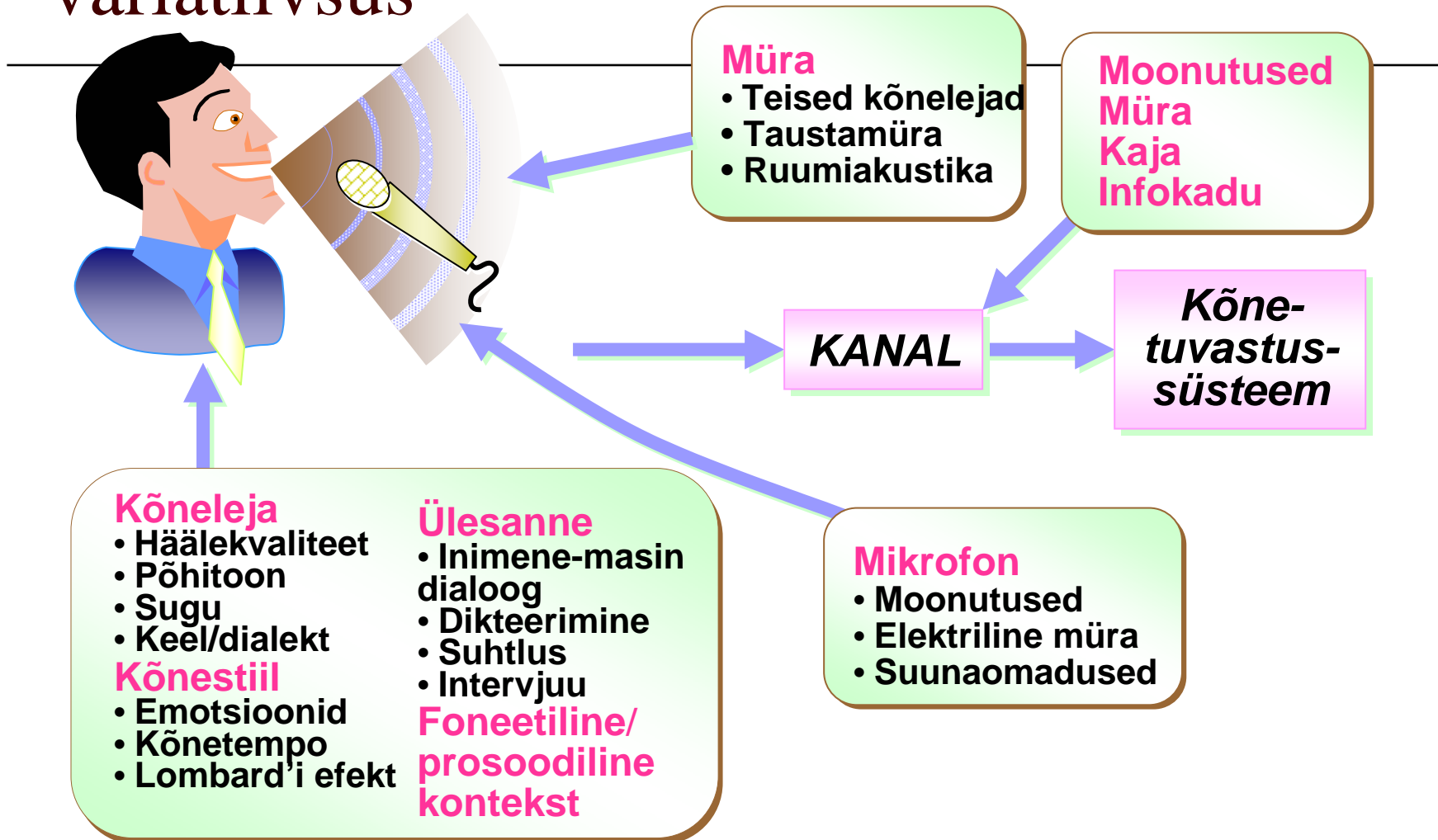
Eesmärgid

Eesti keele foneetilisteks ja kõnetehnoloogilisteks uuringuteks ning arendustöödeks vajalike kõnekorpuste salvestamine, digitaliseerimine, märgendamine ja arhiveerimine, samuti ühtse tehnoloogilise keskkonna loomine erinevate andmebaaside haldamiseks ja efektiivseks kasutamiseks.

Projekti alamülesanded:

1. erinevate kõnekorpuste (spontaanne kõne, dialoogid fikseeritud valdkondades, uudiste lugemine, jms) salvestamine;
2. aktsendikorpuse loomine eri emakeelega isikute eesti keele hääldusnäidetest;
3. kõnesalvestuste infrastruktuuri arendamine ja tehnoloogilise keskkonna loomine erinevate kõneandmebaaside salvestamiseks, haldamiseks ja efektiivseks kasutamiseks;
4. kõnesignaali segmenteerimise ja märgendamise ühtsete reeglite väljatöötamine.

Kõnetuvastuse põhiprobleem – kõne suur variatiivsus



Olulisus

Erinevad kõne andmebaasid

- Mitmekesise ja süstematiseeritud kõnematerjali olemasolu võimaldab uurida erinevaid suulise kõne aspekte – erinevad kõnestiilid, temaatilised valdkonnad, akustilised tingimused, jne
- Kõnelejast tingitud variatiivsus – kõnelejast sõltumatu tuvastussüsteemi loomiseks on treenimisel vajalik kasutada paljude inimeste häälalusnäiteid
- Aktsendi korpuse loomine on vajalik aktsendinähtude akustilise analüüsi ja modelleerimise tarvis, ka aktsendiga kõne tuvastuseks vajalike mudelite treenimiseks

Olulisus

Infrastruktuur & tehnoloogiline keskkond

- Mitmed praegu uurimistöös kasutatavad kõnekorpused on salvestatud eri formaatides ja seetõttu on nende paralleelne kasutamine tülikas – vajalik ühtne tehnoloogiline platvorm ja kasutajaliides
- Kõnesalvestused tuleb koguda uurimisülesandest/rakendusest lähtuvalt:
 - Foneetiline uuring → laboratoorne kõne → **helisalvestusstuudio**
 - Automaatne telefoniteenus → salvestused telefonikanalis
 - Teksti dikteerimine → salvestused büroo/kodukeskkonnas
 - Militaarrakendus → salvestused militaarobjektidel
 - Jne

Olulisus

Segmenteerimise ja märgenduse printsiibid:

- Erinevad uurimisülesanded vajavad kõnesignaalide segmenteerimist ja märgendamist mitmetel eri tasanditel
- Puuduvad ühtsed segmenteerimis- ja märgendamisprintsiibid – ühe uurimisrühma poolt kogutud ja valdkonna-spetsiifiliselt märgendatud kõnekorpused on teistele uurijatele sageli kasutatud

Kõnekorpusete kogumise ja segmenteerimise/märgendamise meetoodika ühtlustamine on vajalik, sest:

- kogumine ja töötlemine on kallis ja töömahukas
- erinevate korpusete oluliselt laiem korduvkasutus
- lihtsam evalveerimine ja rahvusvaheline koostöö
- RP projektide tulemused on vabavara
- maksumaksja raha kokkuhoid ja efektiivsem kasutamine

Milleks korpused?

There are two kinds of linguists: armchair linguists and corpus linguists. Charles J. Fillmore (1992)

Ratsionalism

- Modelleerib keelepädevust
- Reeglipõhine keeletöötlus
- Generatiivne grammatika (jm)
- Olekuautomaadid (lõplik automaat)
- Rakendused:
 - Morfoloogia
 - Süntaks
 - Semantika

Empirism

- Modelleerib keelekasutust
- Andmepõhine keeletöötlus
- Tõenäosuslik grammatika
- Tõenäosuslik olekuautomaat (Markovi mudel)
- Tehisnärvivõrgud
- Rakendused:
 - Morfoloogia
 - Semantika
 - Pragmaatika
 - **Kõnetuvastus**

Vt Mare Koit 2005. Ratsionalism ja empirism keeletöötluses: vastasseis või koostöö?

Progress kõnetuvastuses

Roger K Moore (SPECOM'2005):

*“Progress has **not** come about as a result of deep insights into SLP by humans. Improvements have come from:*

- *‘data-driven’ approach*
- *increase in computer power*
- *benchmark testing”*

Chin-Hui Lee (ICSLP'2004):

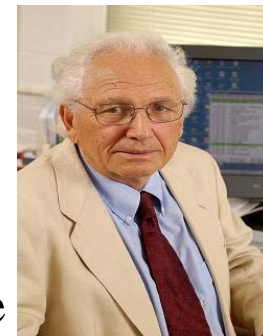
“The field of ASR has enjoyed a fast technological progress in the last three decades, due to the extensive use of:

- *statistical learning algorithms,*
- *the availability of a number of large collections of speech and text examples,*
- *and fast computing machines.”*



Progress kõnetuvastuses

- Kõnetuvastuse põhiline algoritm - *There's no data like more data!*
- **Raimo Bakis** (IBM): “Nende inimestega [keeleteadlastega] pole kõnetuvastuse juures eriti midagi peale hakata” (personaalne vestlus 2003)
- **Frederick Jelinek**: (IBM) “Iga kord kui vallandasin ühe keeleteadlase, paranes süsteemi tuvastusprotsent”
- **Folkloor**: “Kõnetuvastussüsteemi tuvastuskorrektus on pöördvõrdeline selle väljatöötamises osalenud keeleteadlaste arvuga”
- **Suuremahulistele korpustele ja statistilistele algoritmidele ei ole täna võrdväärset alternatiivi!**



Rahvusvaheline praktika

- EAGLES - Expert Advisory Group on Language Engineering Standards <http://www.ilc.cnr.it/EAGLES96/home.html>
- Bavarian Archive of Speech Signals <http://www.phonetik.uni-muenchen.de/Bas/>
- Soome projekt <http://www.csc.fi/kielipankki/puhe/>
- ELRA-ELDA <http://www.elra.info/> <http://www.elda.org/>
- SPEX <http://www.spex.nl/index.php>

Senised tulemused - kõnekorpused

Uudistekorpused:

- ca 300 tunni Eesti Raadio lühiuudiste salvestusi
- digitaliseeritud üle 8000 lk uudistetekste
- arendamisel on tehnoloogiline lahendus ja kasutajaliides uudistekorpuse märgendamiseks
- rakendatakse uudiste automaatse transkribeerimise rakenduse loomiseks

Senised tulemused - kõnekorpused

Aktsendiga kõne korpus:

- on koostatud eestikeelne tekstikorpus – eesti keele häälikud ja häälikuühendid erinevas segmentaalses ja prosoodilises kontekstis, arvesse on võetud sagedasemate aktsendinähtustega, mis eri emakeelega kõnelejate eestikeelses häälduses võivad esineda
- on koostatud register eri emakeele taustaga keelejuhtidest (TÜs ja TLÜs õppivad välistudengid, Eestis töötavad välismaalased), kes esindavad kokku 18 erinevat keelt
- salvestusi alustatakse 2007. aasta lõpus

Senised tulemused - kõnekorpused

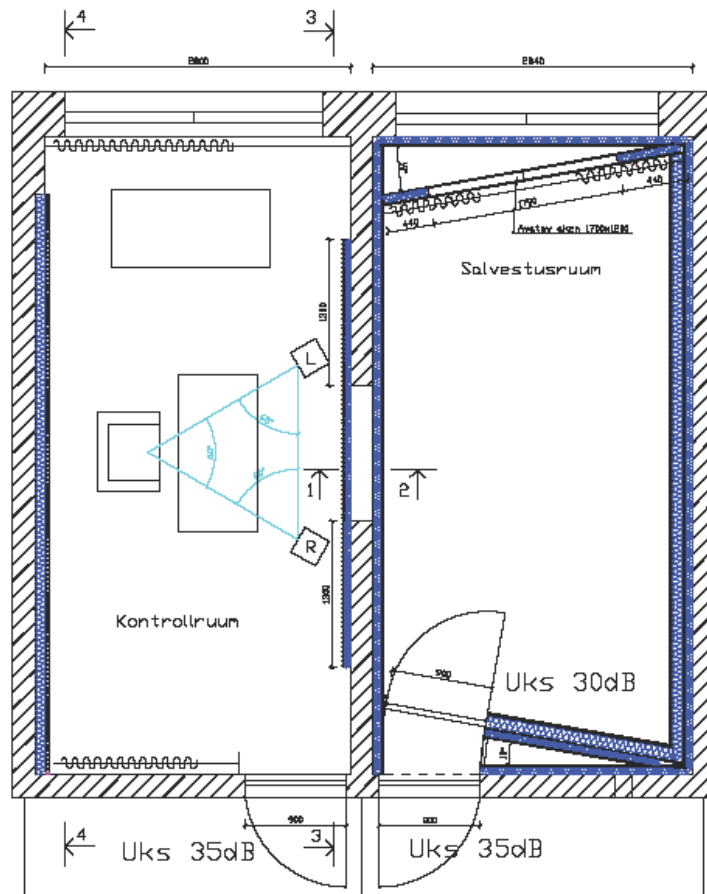
Loengute salvestused:

- on arhiveeritud ca 20 tundi loengukõne salvestusi - põhiliselt TTÜ erinevate ainekursuste salvestused (4 keelejuhti) ja konverentsiettekanded (3 keelejuhti)
- 2007.a lõpuks ca 50 tundi
- kasutatakse loengukõne temporaalse struktuuri uurimiseks (koostöös TÜga) ja spontaanse kõne tuvastuse/automaatse segmenteerimise rakenduse loomiseks (T.Alumäe)

Senised tulemused - kõnesalvestuste infrastruktuuri arendamine

- On välja ehitatud ja sisustatud sobivate akustiliste omadustega kõnesalvestusstuudio (projekteerinud Soome firma Akukon OY)
- Tehnilised võimalused salvestusteks erinevate telefonikanalite kaudu

Salvestusstuudio



Salvestustehnika

- Mikrofonid:
 - AKG
 - Sennheiser
 - Behringer
- Kõlarid:
 - Genelec
 - JBL
- Mikserpult – Mackie Onyx 1640, FireWire I/O-kaart, salvestustarkvara
- Adobe Audition 3.0
- Mobiilsed salvestusvahendid:
 - M-Audio MicroTrack 24/96
 - Edirol R1
 - Digigram VXpocket 440

Senised tulemused - segmenteerimise ja märgendamise reeglid

Salvestuste segmenteerimine:

- automaatne: HMM, NN – force alignment (HTK, Festival)
- käsitsi – Praat, SFS, jt.

Märgenduse tüübid:

- ortograafiline transkriptsioon
- fonemaatiline transkriptsioon (SAMPA)
- foneetiline transkriptsioon (IPA, WORLDBET)
- prosoodiline transkriptsioon (TOBI)
- dialoogiaktid
- kõnelejavoored
- meta-lingvistilised helid: hingamine, naer, köhimine
- kõne/vaikus (mitte-kõne, müra)
- ...

Senised tulemused - segmenteerimise ja märgendamise reeglid

Fonemaatiline transkriptsioon - SAMPA

<http://www.phon.ucl.ac.uk/home/sampa/>

Foneetiline transkriptsioon:

IPA: <http://www.arts.gla.ac.uk/ipa/ipachart.html>

WORLDBET: James L. Hieronymus, *ASCII Phonetic Symbols for the World's Languages: Worldbet*, Technical report, Bell Labs, 1993.

<http://www.ling.ohio-state.edu/~mbeckman/825sp2005/WorldBet/WorldBetAll.pdf>

Senised tulemused - segmenteerimise ja märgendamise reeglid

BABEL-projekti segmenteerimisjuhised

P. Roach *et al.*, 1995. Segmentation and Labeling criteria for the BABEL-database. Project report.

ANDOSL Acoustic-Phonetic Segmentation Criteria

K. Croot, B. Taylor, 1995. Criteria for Acoustic-Phonetic Segmentation and Word Labelling in the Australian National Database of Spoken Language. Speech, Hearing and Language Research Centre, Macquarie University, 1995.

<http://www.shlrc.mq.edu.au/projects/andosl/index.html>

T.Lander, 1997. The CSLU Labeling Guide

<http://cslu.cse.ogi.edu/corpora/docs/labeling.pdf>

M. Vainio, 2001. Artificial Neural Network Based Prosody Models for Finnish Text-to-Speech Synthesis. Publications of the Department of Phonetics, University of Helsinki.

Appendix A: Database labeling criteria and some statistical descriptions of the data.

M. Lennes, S. Ahjoniemi, 2005. Puheaineiston annotaatio eli nimikointi

http://www.helsinki.fi/~lennes/annotation_guide/annotation_guide.html
EKKTT 2007 19.-21.11.2007

Senised tulemused - segmenteerimise ja märgendamise reeglid

Eestikeelsed juhised:

Kõnesignaalide segmenteerimise reeglid.

Koostanud Lya Meister (2005)

Valmimas on juhendi laiendatud versioon:

- Ülevaade erinevatest segmenteerimis- ja märgendusprintsippiidest
- Häälüklasside akustilised tunnused
- Näited erinevatest hääldusvariantidest
- Näited aktsendiga kõnest

Edasine töö, lõpptulemused

Uudistekorpused – ca 300 tundi Eesti Raadio lühiuudiste salvestusi koos ortograafilise märgendusega

Aktsendikorpused – ca 30 eri keeletaustaga eestikeelse kõne salvestused:

- Vene ja soome keel – 50 kõnelejat, a' 15-20 minutit
- EL “suuremad” keeled (inglise, saksa, prantsuse, hispaania) – 20 kõnelejat
- EL “väiksemad” keeled – 4-6 kõnelejat
- Muud keeled – 1-2 kõnelejat

Loengukõne korpused – ca 200 tundi loengute ja konverentsiettekannete salvestusi 15-30 kõnelejalt

Edasine töö, lõpptulemused

(?) **SPEECON korpus** – kõnesalvestused erinevate seadmete hääljuhtimiseks:

- 4 keskkonda – kodu, büroo, avalik ruum, auto
- 4 mikrofoni erineval kaugusel kõnelejast
- kuni 200 keelejuhti
- kuni 1 tund kõnematerjali igalt keelejuhilt

(?) SpeechDat Car

(?) Kõneleja verifitseerimise korpus

(?) Mürakorpus

(?) Noortekõne korpus

(?) Multi-modaalsed korpused

(?) ...

Tegijad, rahastamine

Põhitäitjad:

Einar Meister

Lya Meister

Lepingulised töötajad (uudiste skaneerimine ja märgendus, loengusalvestused, jms)

Finantseerimine:

2006 – 660 000 EEK

2007 – 400 000 EEK

Salvestusruum:

Projekt: 100 000 EEK (Akukon OY)

Ehitus: 400 000 EEK (Hooldusteenus OÜ)

Aparatuur: 100 000 EEK