

# Eestikeelse kõnetuvastuse meetodite uurimine ja arendamine

Tanel Alumäe, Toomas Kirt

Küberneetika Instituut  
Tallinna Tehnikaülikool

Riikliku programmi "Eesti keele keeletehnoloogiline tugi (2006-2010)"  
konverents, 2007

- 1 Sissejuhatus
- 2 2007. a senised tulemused
  - Statistilise keelemudeli adapteerimine
  - Autosegmenteerija
  - Väldete modelleerimine
- 3 Publikatsioonid
- 4 Tulevane töö
- 5 Demo

# Sissejuhatus

## Kõnetuvastus

Kõnetuvastuse käigus konverteeritakse kõnesignaali kõnes esinenud sõnade jadaks.

## Kõnetuvastuse rakendusvaldkonnad

- Suulisel kommunikatsioonil baseeruvad kasutaja-sõbralikud liidesed
- Automatiseeritud dikteerimine
- Suurte kõnekogumite indekseerimine, kõneandmete otsing

## Eesmärgid

Projekti eesmärgiks on eesti keelele sobivate kõnetuvastuse meetodite uurimine, arendamine ja testimine ning erinevate tuvastussüsteemi prototüüpide loomine.

## Sisuline töö

- Seni oleme keskendunud eesti keele spetsiifilistele probleemidele
  - ▶ Eesti keele aglutinatiivsusest ja flekteeruvusest tingitud probleemid statistilise keelemudeli loomisel ja rakendamisel
  - ▶ Häälussõnastiku genereerimine
  - ▶ Kolmevärtelise kestussüsteemi käsitus akustilisel modelleerimisel
- Prototüüpide loomine (dikteerimine, automaatne transkribeerimine, dialoogisüsteemid)

# 2007. a senised tulemused

# Statistilise keelemudeli adapteerimine

# Statistilise keelemudeli adapteerimine

## Statistiline keelemudel

Keelemudelit kasutatakse kõnetuvastuses sõnade aprioorse kontekstuaalsete tõenäosuste arvutamiseks. Tüüpiline keelemudel on trigramm-mudel, kus sõna tõenäosus arvutatakse kahe eelneva sõna põhjal.

## Keelemudeli adapteerimine

- Tavaliselt kasutatakse üldist keelemudelit, mis peaks võrdselt hästi töötama kõigi jututeemade puhul
- Kõne keskendub aga tavaliselt mingile teemale
  - ▶ näit uudiste transkribeerimine: lood sise- ja välispoliitikast, majandusest, kultuurist, spordist
  - ▶ igal teemal on küllaltki spetsiifiline sõnavara
- **Keelemudeli adapteerimine:** teades paari lauset antud teemast, kohanda keelemudeli tõenäosusi nii, et teema-spetsiifilised sõnad ja sõnakombinatsioonid saaks suurema aprioorse tõenäosuse

# Statistilise keelemudeli adapteerimine

## Tüüpiline lähenemine

Vaadeldakse, millised sõnad “seemnelausetes” esinevad; seejärel leitakse dokumendikorpuse statistika põhjal sellised sõnad, mis esinevad tihti samades dokumentides. Nende sõnade aprioorseid tõenäosusi suurendatakse ja teiste omi vähendatakse.

## Probleemid

- Eesti keel on aglutinatiivne ja flekteeruv, palju on liitsõnu. Liitsõnu tekib kogu aeg juurde.
- Sellepärast kasutatakse kõnetuvastuses keele põhiühikutena morfeeme
  - ▶ Morfeemide abil saadakse suur keele **katvus** juba u 60 000 ühiku juures
- Kas morfeeme saab ka kasutada keelemudeli adapteerimiseks
  - ▶ Teisisõnu, kas morfeemid kannavad piisavalt semantilist sisu?

# Statistilise keelemudeli adapteerimine

## Meetodi kirjeldus

Proovisime järgmist lähenemist:

- Kasutada nn *varjatud semantika analüüsi* (*latent semantic analysis* – LSA) dokumentide semantilise läheduse arvutamiseks
- Eksperimenteerida erinevate ühikutega (sõnad, lemmad, morfeemid) semantilise läheduse kujutamiseks
- Adapteerimine:
  - ▶ Nn seemnelausete põhjal leida dokumendikorpusest sellele teemale lähedaseimad dokumendid
  - ▶ Leitud dokumentides esinevate morfeemide statistika abil kohandada keelemudeli tõenäosusi

# Statistilise keelemudeli adapteerimine

## Ekspereimendid

- Kõnetuvastusülesanne: ER täistunnised lühiuudised, segmenteeritud uudislugudeks
- Ekspereimendi kirjeldus
  - 1 Tuvastuse esimeses faasis kasutasime üldist keelemudelit tuvastushüpoteeside saamiseks
  - 2 Igast uudisloost tuvastatud teksti kasutasime semantiliselt lähedaste dokumentide leidmiseks (poole miljoni ajaleheartikli hulgast)
  - 3 Leitud dokumentide põhjal adapteerisime keelemudeli
  - 4 Tuvastuse teises faasis kasutasime adapteeritud keelemudelit
  - 5 Võrdlesime esimese faasi ja teise faasi tuvastusväljundi kvaliteeti

Adapteerimine	LER, %
–	7.1
Sõna-põhine	6.7 (-6%)
Lemma-põhine	6.6 (-7%)
Morfeemi-põhine	<b>6.4 (-10%)</b>

# Autosegmenteerija

# Autosegmenteerija

## Autosegmenteerija

Autosegmenteerija on tarkvara, millega saab segmenteerida eestikeelset kõnet sõnadeks ja häälikuteks, kasutades Markovi peitmudelitel põhinevaid kõnetuvastuse akustilisi mudeleid. Segmenteerimiseks peab kõne olema juba sõnade tasemel transkribeeritud, autosegmenteerimise käigus leitakse automaatselt sõnade ja sõnaosade olevate häälikute piirid.

## Omadused

- Automaatne häälduse leidmine sõna ortograafiast.
- Automaatne täidetud pauside tuvastamine: spontaanse kõne transkriptsioonis ei pea eraldi näitama täidetud pause, nagu köhatused, kõhklushäälitsused, mürad jms. Autosegmenteerija täidab need ise sobivate ühikutega.
- Segmenteerimistulemuste väljastamine Praat TextGrid formaadis: TextGrid formaadis failis on nii sõna-tasemel kui ka hääliku-tasemel segmenteerimispiirid.

# Autosegmenteerija

Demo

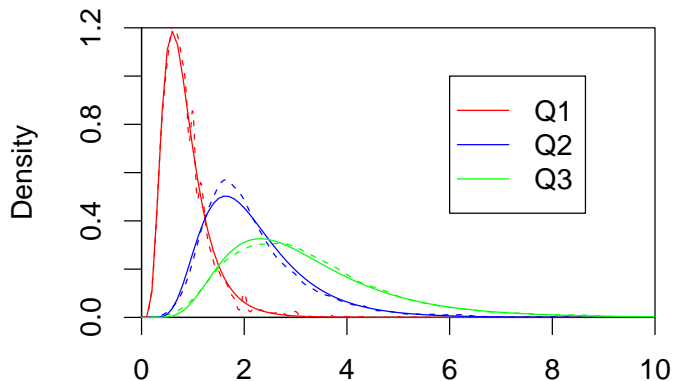
# Väldete modelleerimine

## Taust

- Kuna välde on suprasegmentaalne nähtus, ei saa välteid modelleerida hääliku tasemel Markovi peitmudelitega
- Väldete tajumisel on oluline esimese rõhulise silbi ja järgneva rõhuta silbi kestuste suhe
  - ▶ Täpsemalt: rõhulise silbi tuuma+kooda ja järgneva silbi tuuma kestuse suhe
- Eesmärk: arvutada tuvastuse käigus silpide kestuste suhe ja kasutada seda ühe informatsiooniallikana sõna akustilise tõenäosuse arvutamisel

# Väldete modelleerimine

## Eksperimendid



Duration ratio



# Väldete modelleerimine

## Arutelu

Allikas	Q1	Q2	Q3
Lehiste 1960	0.7	1.5	2.0
Liiv 1961	0.7	1.6	2.6
Eek 1974	0.7	2.0	3.9
Krull 1991	0.5-0.7	1.2-2.1	2.2-2.9
<b>Autosegmenteerija, 50%</b>	0.6-1.0	1.5-2.6	2.1-4.0

Probleemid:

- Suur varieeruvus, teise ja kolmanda välte suur kattuvus
- Morfoloogilise analüsaatori vead kolmanda välte identifitseerimisel
- Autosegmenteerija vead sõna lõpu määramisel

# Publikatsioonid

- 1 Alumäe, T., Kirt, T. *Lemmatized latent semantic model for language model adaptation of highly inflected languages*. Baltic HLT Conference 2007, Kaunas. Ilmumas.
- 2 Alumäe, T., Kirt, T. *LSA-based language model adaptation for highly inflected languages*. Proceedings of Interspeech 2007, Antwerp, Belgium, pp. 2357-2360.
- 3 Alumäe, T. *Automatic compound word reconstruction for speech recognition of compounding languages*. Proceedings of NODALIDA 2007, Tartu, Estonia, pp. 5-12.
- 4 Alumäe, T. *Methods for Estonian large vocabulary speech recognition*. Ph.D. Thesis, Tallinn University of Technology, TUT Press, 2006.
- 5 Treumuth, M., Alumäe, T., Meister, E. *A natural language interface to a theater information database*. Proceedings of IS-LTC 2006, Ljubljana, Slovenia, pp. 27 - 30.
- 6 Alumäe, T. *Sentence-adapted factored language model for transcribing Estonian speech*. In Proceedings of ICASSP 2006. Toulouse, France, vol. 1, pp. 429 – 432.

# Tulevane töö

- $N$ -gramm-mudelist pikemate süntaktiliste seoste modelleerimine keelemudelis
- Spontaanse kõne tuvastusega seotud probleemid
- Väldete modelleerimise edasine uurimine

# Demo

