

Kõnetuvastusest Eestis ja CNRS/LIMSI laboris Prantsusmaal

Tanel Alumäe

Küberneetika Instituut
Tallinna Tehnikaülikool

Riikliku programmi "Eesti keele keeletehnoloogiline tugi (2006-2010)"
konverents, 2009

1 Sissejuhatus

2 CNRS/LIMSI

- Quaero projekt
- ESTER II
- Minu töö LIMSIS

3 EKKTT kõnetuvastuse projektist

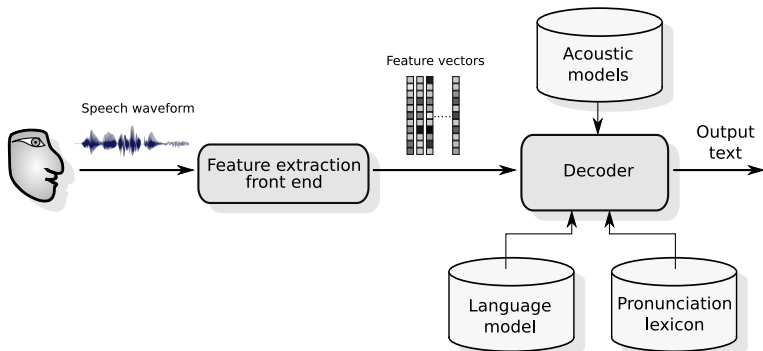
Sissejuhatus

Kõnetuvastus

Kõnetuvastuse käigus konverteeritakse kõnesignaali kõnes esinenud sõnade jadaks.

Rakendused:

- Kirjade jm dokumentide dikteerimine
 - ▶ Reaalajanõue, võimalik adapteerida konkreetsele kasutajale
 - ▶ Kasutaja kooperatiivne, planeeritud kõne
- Kõnesalvestuste automaatne transkribeerimine
 - ▶ Reaalajanõue puudub
 - ▶ Planeeritud või spontaanne kõne
 - ▶ Tüüpiliselt mitu erinevat kõnelejat
- Inimene-arvuti dialoog
- Masintõlge kõnest



CNRS/LIMSI



- *Le Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur* – Mehaanika ja inseneriteaduste laboratoorium, CNRS (Centre national de la recherche scientifique) alluvuses
- 120 töötajat, 60 doktoranti
- Uurimisrühmad:
 - ▶ Audio & Acoustic
 - ▶ Unsteady Aerodynamics : Turbulence and Control
 - ▶ Architecture and Models for Interaction
 - ▶ Convection and Rotation : Instabilities and Turbulence
 - ▶ Language, Information and Representations
 - ▶ Perception & Cognition
 - ▶ Virtual and Augmented Reality
 - ▶ **Spoken Language Processing**
 - ▶ Solid-Fluid Transfers

Uurimisteemad:

- Akustilis-foneetilised mudelid
- Keelemudelid
- Ringhäälingu-uudiste transkribeerimine
- Vestluskõne transkribeerimine
- Kõnelejatuvastus
- Emotsioonituvastus
- Keeletuvastus
- Audio indekseerimine
- Teema jälgimine (topic tracking)
- Kõne semantiline analüüs
- Dialoogisüsteemid
- Masintõlge

Quaero projekt



- Suur Euroopa R&D programm
- Kogu eelarve üle 200 miljoni euro
- Põhilised uurimissuunad: automaatne info eraldamine multilinguaalsetest multimeedia dokumentidest (kirjalikud tekstid, kõne, video, muusika, pildid), ning selle info klassifitseerimine, analüüs ja kasutamine
- Partnerid: Thomson, Exalead, Bertin Technologies, Jouve, Grass Valley GmbH, Vecsys, Vecsys Research, LTU Technologie, Synapse Développement
- Uurimisasutused: CNRS/LIMSI, INRIA, IRCAM, RWTH Aachen, University of Karlsruhe, IRIT, Clips Imag, GET, INRA
- Rakendused:
 - ▶ multimeedia indekseerimise ja otsinguvahendid professionaalseks ja tavakasutuseks
 - ▶ professionaalsed vahendid multimeedia dokumentide tootmiseks, halduseks ja levitamiseks
 - ▶ digitaliseeritud kultuuripärandile (näit. audiovisuaalsed arhiivid) parema juurdepääsu võimaldamine

Reaalsed hetkel uuritavad teemad:

- Kõnetuvastus, keele identifitseerimine kõnest, kõne segmenteerimine kõnelaja järgi, Q&A süsteemid, masintõlge, aktsendi klassifitseerimine
- Näo/isiku leidmine ja tuvastus (fotolt ja videolt)
- Video segmenteerimine, sündmuse identifitseerimine videos
- Videote ja piltide sõrmejäljestamine ja identifitseerimine
- Muusika klassifitseerimine, muusika otsing sarnasuse põhjal, muusika segmenteerimine ja kokkuvõtte genereerimine
- Nimisõnafraaside ja terminite leidmine, ontoloogiate loomine, semantiline annotatsioon
- Dokumendiotsinguga seotud teemad, dokumentide struktureerimine

Kõnetuvastusest QUAERO projektis:

- Partnerid: LIMSI, RWTH Aachen, Karlsruhe ülikool
- Praegu ainult inglise, saksa ja prantsuse keelele
- Peatselt lisandub hispaania ja portugali keel
- Tegeletakse rootsi, kreeka ja soome keelega
- Eesmärgiks luua tuvastustehnoloogia kõigile EL keeltele, lisaks mandarini ja araabia, esialgu ainult “vanadele” EL keeltele
- Suur osatähtsus statistilistel meetoditel, vähe lingvistilisi meetodeid

ESTER II

- Prantsusekeelse kõne transkribeerimise võistlus/hindamine
- Osalevad kõnetuvastusega tegelevad uurimisrühmad
- Inspireeritud NIST Rich Transcription hindamisest USAs
- Osalemine vabatahtlik
- Transkribeeritav kõne: prantsuskeelsete raadiote uudistesaadetest
- Eesmärk: ressursside loomine, uuringute ja suhtluse edendamine, süsteemide hindamine

ESTER II: tuvastusülesanne

- Osavõtjatele jagatakse nn arendusülesanne (*development set*) oma süsteemi tuunimiseks – kõne koos täpse märgendusega, mis peaks olema lõppülesandele üsna sarnane
 - ▶ Selle aasta arendusülesandes nelja erineva regiooni raadiote uudistesaaated: France Inter (Prantsusmaa), RFI (Prantsusmaa, välisuudised), Africa 1 (erinevad prantsusekeelsed Aafrika maad), TVME (Maroko). Uudistesaaated pikkusega 10 minutit kuni 1 tund, sisaldavad muusikat, telefoniintervjuusid, lõike spordireportaažidest jms
- Treeningmaterjali ei anta, see peab endal olemas olema
- Lõppülesande (nn *test/evaluation set*) kõne peab töötleva valmis süsteemiga “pimedalt” ja saadud märgendused korraldajale tagasi saatma. Salvestuste metaandmeid (raadio, kuupäev, kellaaeg) võis kasutada
- Korraldaja leiab iga osavõtja märgenduse kvaliteedi ja teeb teatavaks tulemused

LIMSI süsteem ESTER II jaoks

Keelemudel

- Statistilise keelemudeli treenimiseks kokku 1.76 G sõna
 - ▶ Ajalehetekstid
 - ▶ Tekstid uudisteagentuuridest
 - ▶ Tekstid veebist
 - ▶ TV-uudiste subtiitrid
 - ▶ Juba transkribeeritud raadiouudiste tekstid (u 8 M sõna)
- 200K sõnaga keelemudel, optimeeritud Prantuse, aafrika ja maroko andmete põhjal
- Keelemudeliks 4-gramm mudel, saadud erinevate treeningtekstide põhjal treenitud mudelite interpolatsioonina
- Prantsuse, aafrika ja maroko uudiste jaoks eraldi keelemudelid

Perplexity:

	Prantuse	Aafrika	Maroko
Üldine keelemudel	71.1	75.2	78.1
Spetsiifiline mudel	71.1	72.4	72.7

LIMSI süsteem ESTER II jaoks

Akustilised mudelid

Dialekti-spetsiifilised akustilised mudelid:

- Üldine prantsuse mudel, treenitud 550h kõne põhjal (k.a. dialektid)
- Telefonikõne mudel
- Prantsuse aafrika dialekti mudel, s.t. üldine prantsuse mudel, adapteeritud 65h aafrika kõne põhjal
- Irgale liigile mehe ja naisespetsiifilised mudelid

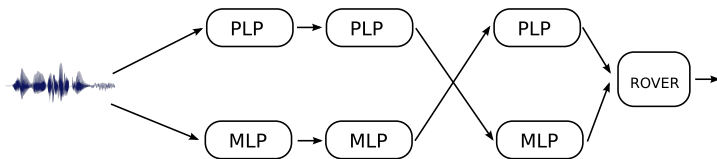
Vigade protsent aafrika kõne puhul erinevate adapteerimisandmete puhul:

	Tunde	WER
Prantsuse	630	25.1
Aafrika	630+56	23.4
Aafrika + pr. kolooniad	630+56+9	23.3
Aafrika + pr. kolooniad + maroko	630+56+9+16	23.7

LIMSI süsteem ESTER II jaoks

Tuvastusstrateegia

- 1 Helisalvestus jagatakse kõnega ja kõneta (vaikus, muusika) lõikudeks
- 2 Kõnega lõigud grupeeritakse kõneleja järgi (*speaker diarization*)
- 3 2 paralleelset tuvastussüsteemi (erinevatel tunnustel põhinevad), mõlemas 3 tuvastus- ja järgnevat akustiliste mudelite adaptsoonifaasi, rist-adaptsoon
- 4 Paralleelsete süsteemide tulemused kombineeritakse ROVER tehnikaga



LIMSI süsteem ESTER II jaoks

Tulemused

Päritolu	WER %
Aafrika	16.8
Maroko	14.7
Prantsuse (Inter)	12.0
Prantsuse (RFI)	11.6
Kokku	14.2

20070619_0730_0745_africa1.wav: WER 13.1%

REF: en république démocratique du congo des manifestations sont prévues aujourd'hui à kinshasa par des organisations professionnelles des médias pour protester contre l' assassinat de serge **** MAHESHÉ et la tentative de meurtre d' anne marie ** ***** KALANGA

HYP: en république démocratique du congo des manifestations sont prévues aujourd'hui à kinshasa par des organisations professionnelles des médias pour protester contre l' assassinat de serge MAIS CHEZ et la tentative de meurtre d' anne marie TA LANGUE A

REF: la présentatrice d' une émission à la radio télévision nationale congolaise et son jeune frère ont été atteints par des balles tirées par trois hommes qui tentaient de pénétrer à leur domicile dans un quartier de l' ouest de la capitale

HYP: la présentatrice d' une émission à la radio télévision nationale congolaise et son jeune frère ont été atteints par des balles tirées par trois hommes qui tentaient de pénétrer à leur domicile dans un quartier de l' ouest de la capitale

Minu töö LIMSIS

Keelemudeli adapteerimine “värskete” andmetega

- Probleem: transkribeerida raadio-uudiseid, olemas on ka samast ajast pärit ajaleheartiklid. Kuidas adapteerida olemasolevat keelemudelit nii, et uudseid või hooajaliselt olulisi sõnu ja sõnakombinatsioone tuvastataks paremini?
- Eriti oluline uute uudistesse kerkivate nimede puhul
- Andmed: ESTER2 uudistesalvestused, prantsuse Google News (u. 400 artiklid päevas) adapteerimiseks

Keelemudeli adapteerimine “värskete” andmetega

- Kõigepealt proovisime adapteerida kõikide artiklitega salvestuse kuupäeva ümbrusest – ei andnud mingit tulemust
- Seejärel filtreerisime dokumente: jätsime alles ainult need artiklid, mis olid tuvastatavale uudisele sarnased
 - ▶ Kasutasime esialgseid tuvastushüpotese ning leidsime need artiklid, milles esines suhteliselt palju salvestuses esinevaid olulisi sõnu
- See andis 10% keelemudeli *perplexity* paranemise
- Kõige parema tulemuse andis +/- 14 päevase vahemiku kasutamine salvestuse kuupäeva ümbrusest
- Tuleviku tekstid sama olulised või isegi olulisemad kui mineviku tekstid
- Tuvastusvigade arv: 13.78% → 13.54%

EKKTT kõnetuvastuse projektist

EKKTT kõnetuvastuse projektist

ER uudistekorpuse automaatne märgendamine

- Uudistekorpus sisaldab ca 300 tundi 10 erineva raadiodiktori loetud Eesti Raadio uudiseid ajavahemikust 29.11.2005 - 29.05.2006
- Uudistetekstid esitati paber kandjal (ca 10000 lk), need skanneeriti
- Paljusid tekste kasutati päeva jooksul korduvalt, muudetult ja lühendatult
- Kuidas ühitada signaalid ja tekstid?

EKKTT kõnetuvastuse projektist

ER uudistekorpuse automaatne märgendamine

- Loodi automaatne uudiste märgendamissüsteem
- Enne mingi päeva uudistesaadete märgendamist luuakse sellele päevale kohandatud keelemudel
- Keelemudeli loomiseks kasutatakse sellel päeval kasutatud diktoritekste
- Keelemudeliks on tavaline trigramm-mudel
- Keelemudelit interpoleeritakse üldise mudeliga, et saavutada mõistlik tulemus ka sellise kõne puhul, mille tekste meil ei ole
- Automaatsed transkribeeritud konverteeritakse automaatselt Transcriberi formaati, ning need vaadatakse märgendajate poolt üle, kes parandavad vead

EKKTT kõnetuvastuse projektist

ER uudistekorpuse automaatne märgendamine

Näide automaatselt märgendusest

Eesti raadiote jutusaadete automaatse transkriptsioonisüsteemi prototüübi loomine

- Uudiste transkribeerimine pole lõpptarbijale eriti huvitav – uudiseid saab ka ajalehest lugeda!
- Vestlussaadete transkribeerimine oleks palju kasulikum
- Viimasel ajal väga paljud raadiote jutusaated netist järelkuulatavad
- Jutusaadete sisu kohta raske ülevaadet saada, ainsaks võimaluseks saate läbikuulamine
- Vestlussaadete automaatne ja piisavalt kvaliteetne transkribeerimine võimaldaks neid paremini indekseerida ja organiseerida, võimaldaks märksõnapõhist otsingut

Eesti raadiote jutusaadete automaatse transkriptsioonisüsteemi prototüübi loomine

- Selle aasta eesmärk:
 - ▶ Luua olemasolevatel treeningandmetel põhinev transkriptsioonisüsteem
 - ▶ Kasutada automaatset kõnelejapõhist adapteerimist
 - ▶ Hinnata nõnda saavutatavat tuvastuskvaliteeti
 - ▶ Tuvastuskvaliteedi parandamiseks vajalike tööde ja ressursside identifitseerimine
- Platvorm mitmesugustele teistele keele- ja kõnetehnoloogilistele uuringutele:
 - ▶ nimisõnafraaside ja terminite leidmine
 - ▶ transkribeeritud kõne struktureerimine ja organiseerimine (näit. kõnelejate ja teemade järgi)
 - ▶ automaatne kokkuvõtete genereerimine kõnest

Küsimused?