Aalto University

DigiTala

Automatisk taligenkänning som hjälpmedel för att bedöma muntliga språkfärdigheter

Mikko Kurimo

Institutionen för signalbehandling och akustik

Aalto-universitetet

Innehåll

1. Hur automatisk taligenkänning fungerar
2. Taligenkänning för svenska och finska
3. Hur sker automatisk uttalsbedömning?

Mikko Kurimo

Associate professor in **speech and language processing**

Background from machine learning algorithms and pattern recognition systems

PhD 1997 at TKK on speech recognition training algorithms

Research experience in several top speech groups:

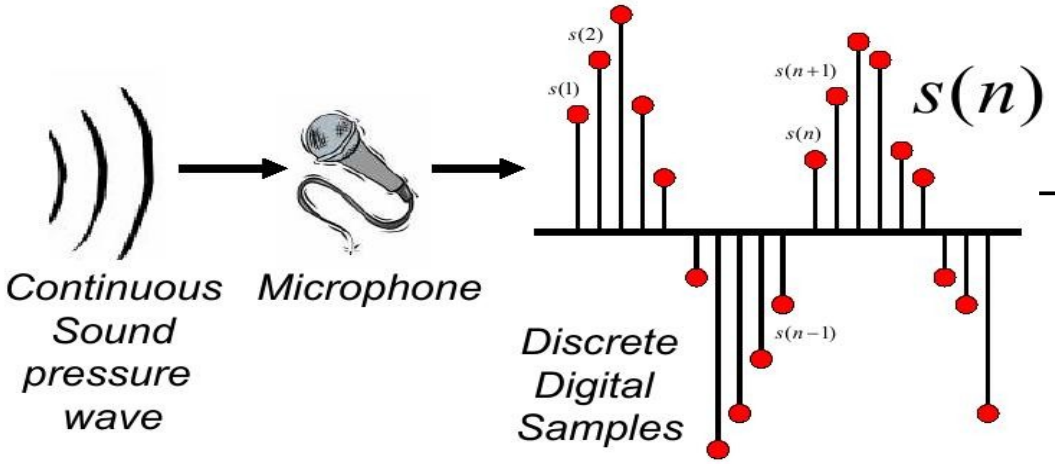
- Research Centers: IDIAP (CH), SRI (USA), ICSI (USA)
- Universities: Edinburgh, Cambridge, Colorado, Nagoya

Head of Aalto **speech recognition research group +** several national and European speech and language projects

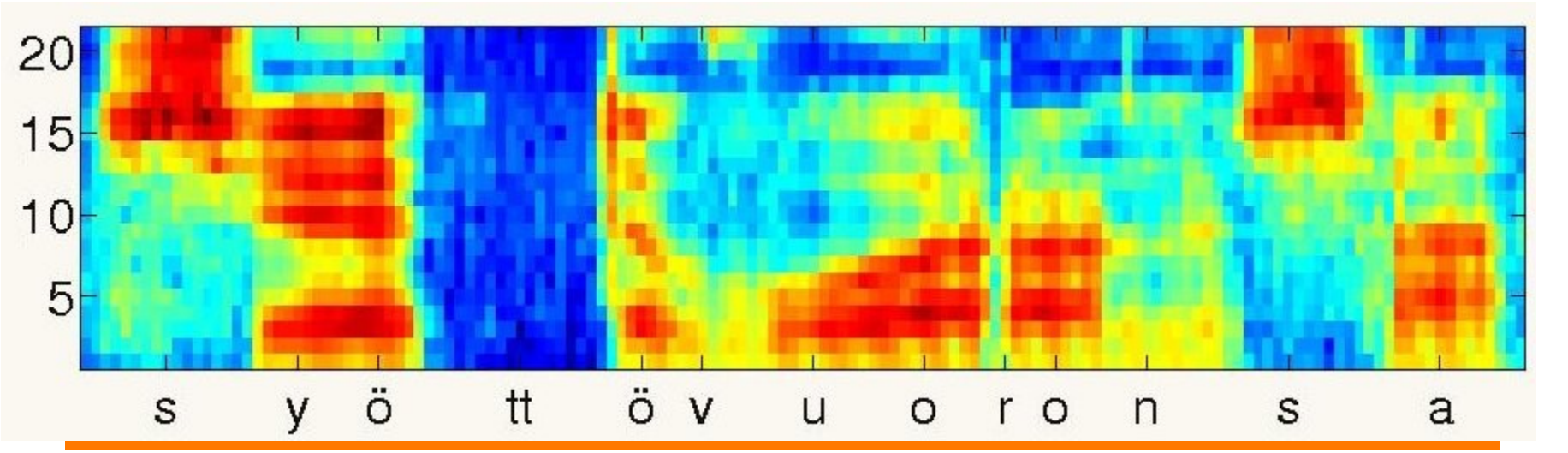
Research topics:

- Speech recognition, language modeling, speaker adaptation, speech synthesis, translation, information retrieval from audio and video

Talsignalen representerad i frekvensspektrumet



21 olika särdrag räknas ut i 10ms intervall från ett effektspektrum



Ljudmodeller

Hur ett ljud uppfattas beror på:

talaren, talhastigheten, talstilen, talsituationen

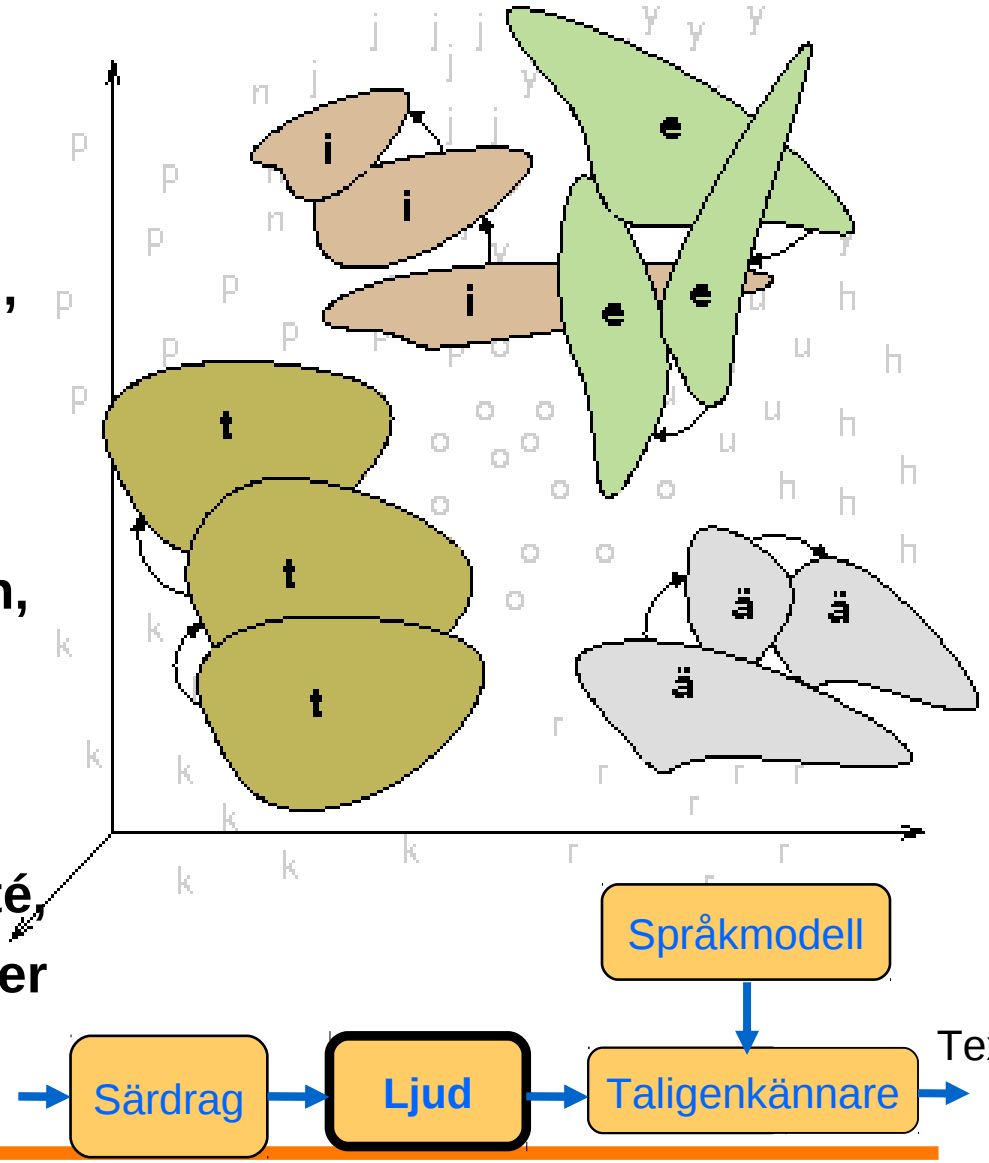
Andra faktorer som påverkar:

bakgrundsljud, eko, mikrofonen, transmissionskanalen

En exakt ljudmodell kräver:

mycket taldata av lämplig kvalité,

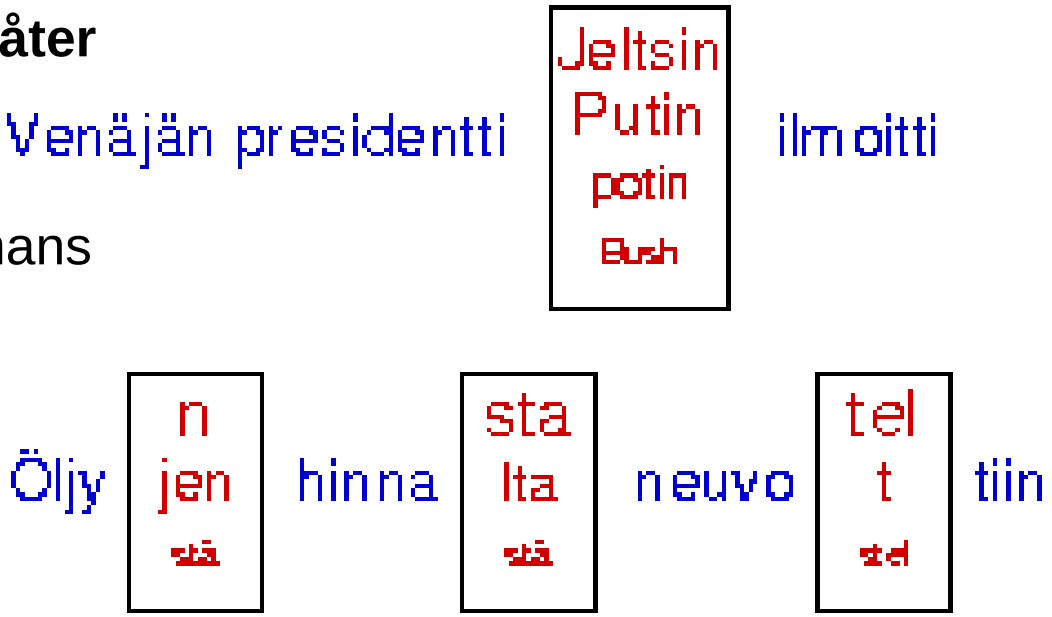
personliga (adapterade) modeller



Språkmodellens uppgift

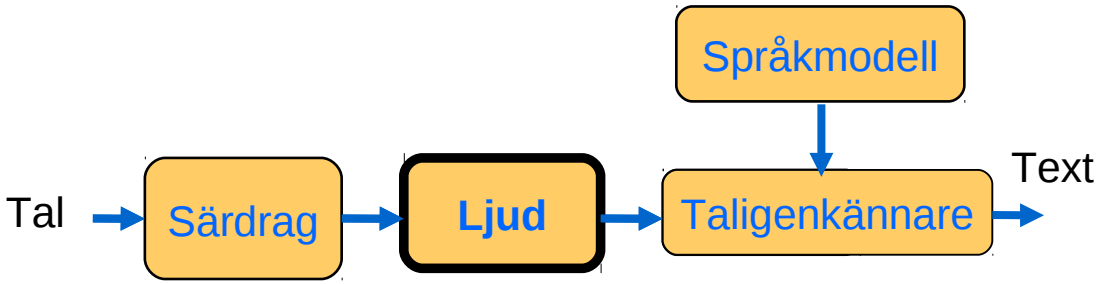
Urskiljer meningar och ord **som låter liknande** och reducerar antalet **alternativ** som undersöks:

- vilka ord förekommer tillsammans
- definierar ord och morfem
- beskriver ordens uttal som ljudsekvenser



Kräver mycket **lämplig data**:

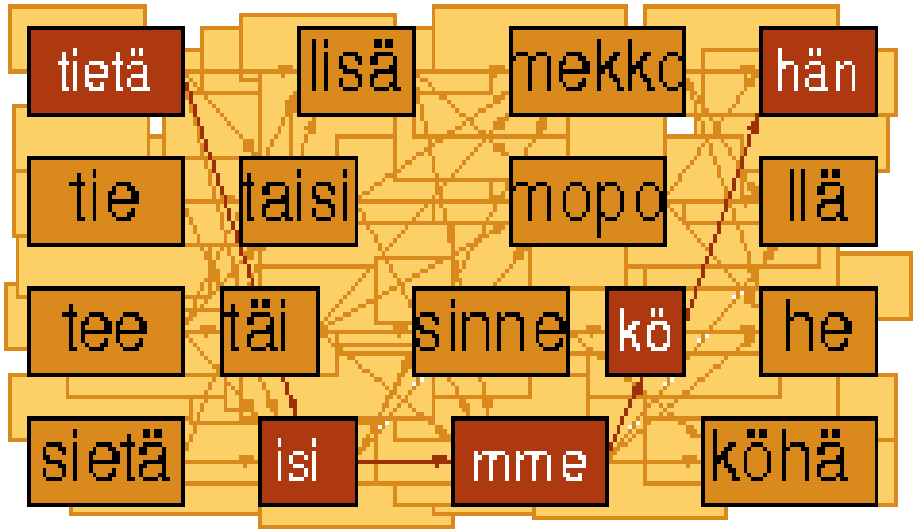
- **skrift- vs. talspråk**
- **planerat vs. spontant**
- **yrkesjargong**

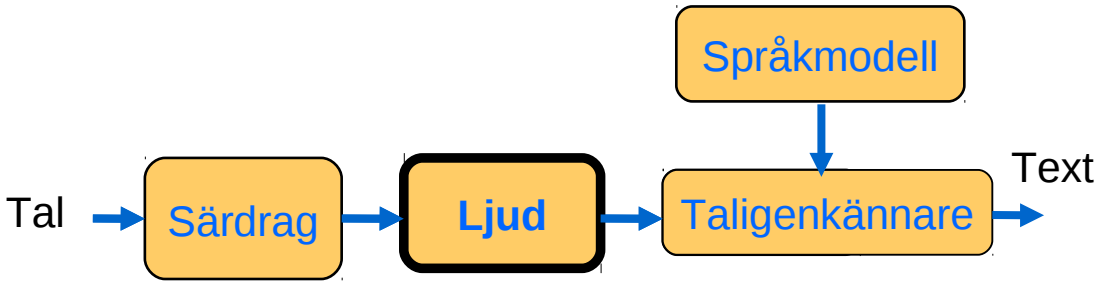


Taligenkännarens uppgift:

Räkna ut den **mest sannolika meningen** med tanke på:

- Inspelade talsignalen
- Ljudmodellen
- Språkmodellen
- Räknekapaciteten





Igenkänningsnogrannheten beror på:

- Igenkännarens träning och testning:
  - materialets storlek och lämplighet
- Inspelning och bakgrundsljud
  - mikrofonen och avståndet
- Talare och talstil
  - byte av talare
  - talarens klarhet och stil
- Språket och talsituationen
  - bok- vs. skriftspråk
  - planerat vs. spontant
  - yrkesjargong

Senaste doktorsavhandlingarna inom forskningsgruppen:

- Audiovisuell indexering och sökning (Ville, 2012)
- Språkmodeller och morfem (SamiV, 2012)
- Diskriminativ träning av ljudmodeller (Janne, 2013)
- Taligenkänning i brus (Ulpu 2016, Heikki 2016)
- Språkmodeller och morfologi (Matti, Stig-Arne, Peter, Teemu 2016)
- Talspråk, låneord och adaptering av språkmodeller (Seppo, Andre 2016)
- Talar-adaptering och talsyntes (Reima)

Andra projekt

- Unsupervised morphological analysis of words
- Speech and language modeling in brains
- Multimodal speaker segmentation and recognition
- Automatic captioning of multimodal data
- Pronunciation evaluation and helping to speak a foreign language

Applikationer inom taligenkänning

1. **Diktering**
2. Uttalsbedömning
3. Indexering
4. Användargränssnitt

1. Bearbetning och behandling av diktat

- **Mål**: En text som beskriver vad som sägs i diktatet. Behövs ingen perfekt transkription, men det väsentliga innehållet ska vara **100% rätt.**
- Automatisk **taligenkänning är ett viktigt hjälpmedel för att göra processen snabbare**, men räcker oftast inte till på egen hand. För brev, rapporter och maskinöversättning borde texten vara felfri. Det sammal gäller för bedömning av språkkunskaper.
- Ett sätt att behandla svårbegripligt tal är att **diktera om på nytt** och spara till en skild inspelning som är sedan lättare för automatisk taligenkänning (vilket medför färre fel och det behövs mindre tid för att rätta texten).

Finsk och svensk taligenkänning

Utmaningarna växer i pilens riktning nedåt

- **Finska**: Tiotals år av erfarenhet vid Aalto, igenkännaren gör inga större fel vid *normaldiktering, WER 20%* även utan talar-adaptering
- **Svenska** (standard rikssvenska): Påbörjat 2015, mycket tal- och textdata och hyfsat resultat, *WER 23%*
- **Finlandssvenska**: Lite data tillgängligt, sökes efter mera, igenkännare är under arbete.
- **Finskspråkigas svenska**: Kan samma system användas som för finlandssvenska? Kan träningsdata fås av YLE?
- **Finskspråkiga skolelevers svenska**: Data har samlats under det här projektet, igenkännare är under arbete.

Varför är skolsvenska svårt för taligenkännaren?

- För tillfället finns bara taligenkännare för rikssvenska, vilket inte motsvarar finskspråkigas uttal
- Skolbokstexter borde vara ok (borde inte vara svårare för rikssvenska språkmodeller!)
- Elevernas uppläsning är inte lika flytande som för svenskspråkiga och de kan göra fel
- Elevernas uttal kan vara ganska långt från nativt uttal och innehålla fel
- Det kan finnas mycket variation i ljudlängder och prosodi vilket medför problem

Observationer från de första taligenkänningsresultaten – våren 2015

- **Stor variation mellan talare** (20 bästa med i testet):
  - Bra: *munkka_2pgz* 56.0% och *munkka_jv5z* 57.5%
  - Dåliga: *olari_4j53* 106.5%, *olari_k38w* 110.4%
- **Bakgrundsljud och andra inspelningssvårigheter** är det största problemet för taligenkänning!

.

WER% = (substitutioner+insertioner+deletioner) / (antalet ord)

Exempel där WER är runt 42% (8 rätt, 6 fel):

när folk frågar mig vad jag gör på fritiden måste jag ofta svara ingenting...

=> när folk frågade nej vajar gör fritt igen måste jag ofta svara ingenting...

De senaste resultaten

- Bakgrundsljud och inspelningsvårigheter minskade vid inspelningarna under hösten 2015, finlandssvenska nu det största problemet
- Igenkänningsnogrannhet (WER%):
  - rikssvenska talare (Aalto/KTH) 23% / 24%*
  - finlandssvensk:
    - talar-/accent-/ingen adaptering 38% / 42% / 47%**
  - gymnasisters skolsvenska: X% / 65% / 65%**
- (* KTH:s publicerade testdata)
- (**uppläsningar ur lärobok)

Applikationer inom taligenkänning

1. Diktering
2. **Uttalsbedömning**
3. Indexering
4. Användargränssnitt

Hur sker automatisk uttalsbedömning? – olika bedömningskriterier:

1. Känner taligenkännaren igen ordet när det uttalas?
2. Hur nära är ordet från igenkänningströskeln?
3. Hur lika är dom olika ljuden jämfört med en modelltalare?
4. Hur lika är prosodin jämfört med en modelltalare?

Hur sker automatisk uttalsbedömning? – olika bedömningskriterier:

1. Känner taligenkännaren igen ordet när det uttalas?
   - nogrannheten inte stor för skolelever, kan känna igen vad som helst, modellerna motsvarar en genomsnittlig nativtalare
2. Hur nära är ordet från igenkänningströskeln?
3. Hur lika är dom olika ljuden jämfört med en modelltalare?
4. Hur lika är prosodin jämfört med en modelltalare?

Hur sker automatisk uttalsbedömning? – olika bedömningskriterier:

1. Känner taligenkännaren igen ordet när det uttalas?
   - nogrannheten inte stor för skolelever, kan känna igen vad som helst, modellerna motsvarar en genomsnittlig nativtalare
2. Hur nära är ordet från igenkänningströskeln?
   - jämförs med ord som har liknande uttal
   - ljudmodellerna kan adapteras till skolelever
   - samma som igenkännarens pålitlighet, är ett rätt eller fel ord helt klart bäst
3. Hur lika är dom olika ljuden jämfört med en modelltalare?
4. Hur lika är prosodin jämfört med en modelltalare?

Hur sker automatisk uttalsbedömning? – olika bedömningskriterier:

1. Känner taligenkännaren igen ordet när det uttalas?
2. Hur nära är ordet från igenkänningströskeln?
3. Hur lika är dom olika ljuden jämfört med en modelltalare?
   - man matchar först rätt ord med talet
   - man mäter hur nära de olika uttalsljuden är från modelltalaren
   - modelltalaren kan väljas (och adapteras) till att vara så lik eleven som möjligt (men med rätt uttal)
   - man kan mäta om ljuduttalen förbättras efter repetition
4. Hur lika är prosodin jämfört med en modelltalare?

Hur sker automatisk uttalsbedömning? – olika bedömningskriterier:

1. Känner taligenkännaren igen ordet när det uttalas?
2. Hur nära är ordet från igenkänningströskeln?
3. Hur lika är dom olika ljuden jämfört med en modelltalare?
4. Hur lika är prosodin jämfört med en modelltalare?

- man matchar först rätt ord med talet
- ljudlängder
- intonation och betoning (F0)

Kontaktuppgifter: Mikko Kurimo

mikko.kurimo@aalto.fi

http://spa.aalto.fi/en/research/research_groups/speech_recognition/demos