



## Automatisk igenkänning av tal och talare

Mats Blomberg

Tal, musik och hörsel  
KTH

GSLT Tal- och taligenkänning M Blomberg 2002-9-9 [ 1 ]



## Automatisk igenkänning av tal

Mats Blomberg

Tal, musik och hörsel  
KTH

GSLT Tal- och taligenkänning M Blomberg 2002-9-9 [ 2 ]



## Översikt - taligenkänning

- Inledning
- Problem
- Akustiska analysmetoder
- Igenkänningstekniker
  - mönstermatchning
  - olinjär tidstjörning (dynamisk programmering)
  - dolda Markovmodeller
  - kunskapsbaserade metoder
  - neurala nät
- Databaser
- Resultat
- Aktuell forskning
- Tillämpningar

GSLT Tal- och taligenkänning M Blomberg 2002-9-9 [ 3 ]



## Varför taligenkänning?

- Naturligt sätt att kommunicera
  - Snabbare inläring
- Effektivare kommunikation
  - Komplexa samband kan uttryckas enkelt i ett språk vi redan kan
- Ersätter tangentbord eller knappsats
  - handdator, telefon, mobiltelefon
- Fungerar i besvärliga miljöer
  - mörker, kyla etc. (dock sämre i buller)
- Händer och syn blir fria för andra uppgifter

GSLT Tal- och taligenkänning M Blomberg 2002-9-9 [ 4 ]



## Tillämpningar

- Handikaphjälpmedel
  - rörelsehindrade
- Telefontjänster
  - intelligenta "telefonsvarare", "e-mail"
  - informationssökning, biljettbokning
- Fria händer
  - diktering
  - styra mobiltelefon
- Studiehjälp
  - tålmodig lärare
  - språkinläring, uttalsundervisning
- Indexering och sökning
  - radio- och TV-program

GSLT Tal- och taligenkänning M Blomberg 2002-9-9 [ 5 ]



## Klassificering av igenkänningsmetoder

- **Vad** känner man igen?
  - enstaka ord, kommandon, diktering, dialog, spontant tal
- **Vem** känner man igen?
  - en talare: talarberoende, -adaptiv
  - alla talare talarberoende
- **Hur** känner man igen?
  - kunskapsbaserade metoder
    - expertsystem med fonetisk kunskap
    - igenkänning via syntes
  - inlärande metoder (statistiskt baserade)
    - dynamisk programmering (DP)
    - dolda Markovmodeller (HMM, Hidden Markov Models)
    - artificiella neuronnät (ANN, Artificial Neural Networks)

GSLT Tal- och taligenkänning M Blomberg 2002-9-9 [ 6 ]

**Svårighet: tal kontra skrift**

- I fonetisk transkription eller vanlig ortografi beskrivs talet med avgränsade, diskreta enheter
- Talet har ett kontinuerligt förlopp pga artikulatorernas mekaniska tröghet
  - Koartikulation:
    - fonem uttalar olika i olika kontext (jfr /s/ i "visir" och "ozon")
  - Reduktion:
    - I snabbt tal och i obetonade stavelser uppnås inte det avsedda uttalet
    - Fonem och stavelser kan falla bort ("bafatt", "Sötälje", "dnasba")

GSLT Tal- och talteknologi M Blomberg 2002-9-9 | 7 |

**Svårigheter - stor variabilitet**

- Mellan talare**
  - Ålder
  - Kön
  - Anatomi
  - Dialekt
- Inom en talare**
  - Stress
  - Sinnesstämning
  - Hälsotillstånd
  - Formellt / Spontan
    - Reduktioner
    - Minsta ansträngning
- Omgivning**
  - Additivt brus
  - Rumsakustik
- Mikrofon, Telefon**
  - Bandbredd
  - Störningar
    - brus
    - frekvensgång
    - transienter
    - clickar
- Lyssnare**
  - Ålder
  - Modersmål
  - Hörsel
  - Bekant / Okänd
  - Människa / Maskin

GSLT Tal- och talteknologi M Blomberg 2002-9-9 | 8 |

**Överlappning mellan vokaler för olika talare**

- Spridning för de två lägsta resonans-frekvenserna (F1 och F2) hos isolerade svenska vokaler uttalade av manliga och kvinnliga talare (G Fant)
- Främre vokaler har väsentlig överlappning.

GSLT Tal- och talteknologi M Blomberg 2002-9-9 | 9 |

**Rumsakustik**

Samma inspelade yttrande uppspelat i två olika rum

Ekofritt rum

Föreläsningssal (KTH:E5)  
Mikrofonavstånd ~3 m

"Nu är det stjälk"

GSLT Tal- och talteknologi M Blomberg 2002-9-9 | 10 |

**Tal i brus**

Inspelat i bil, hastighet 90 km/t.  
Riktad mikrofon i instrumentpanelen

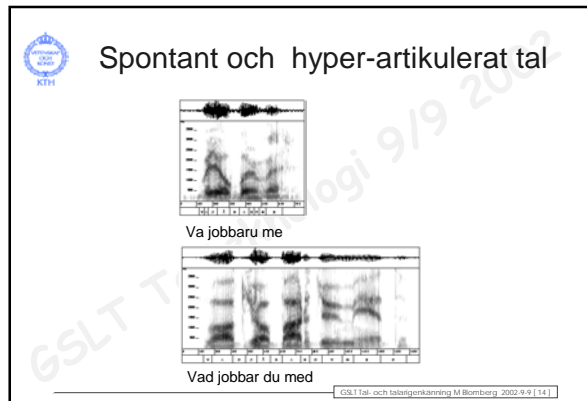
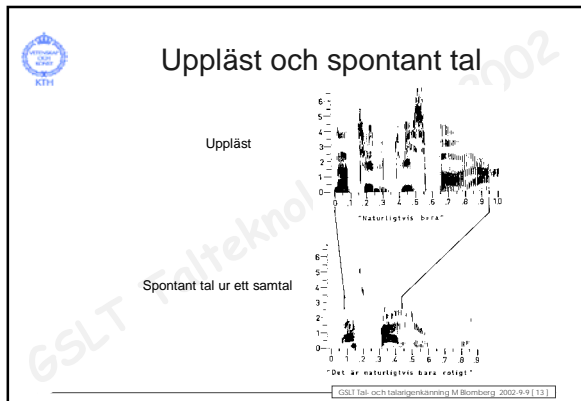
Yttrande: "Inga"

GSLT Tal- och talteknologi M Blomberg 2002-9-9 | 11 |

**Talspråk: extra svårigheter**

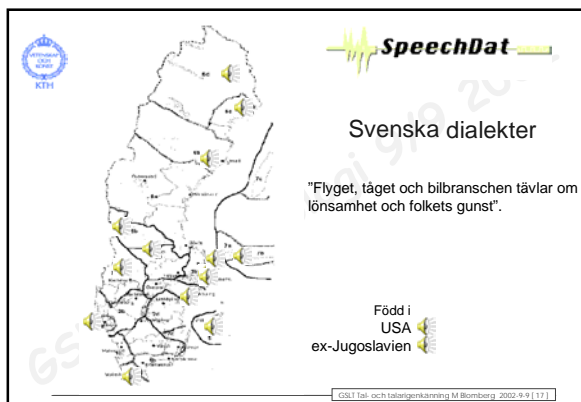
- Uttalsreduktioner
  - ofullständigt uttalade ord
- Ikke-grammatiska meningar
- Stakningar
  - omstarter, instopp, strykningar
- Extralingvistiska ljud
  - läppsmack, andning, tvekljud
- Störningar
  - omgivningsljud, teknisk distorsion

GSLT Tal- och talteknologi M Blomberg 2002-9-9 | 12 |



- Träning**
- För att ett system ska känna igen tal oberoende av talare och miljö behövs kvantitativa mått på denna variabilitet.
  - Ett stort träningsmaterial krävs för att uppskatta dessa
  - Automatiska träningsmetoder nödvändiga
- GSLT Tal- och taligenkänning M Blomberg 2002-9-9 | 15 |

- EU-projektet** 
- Inspelat tal över telefonnätet för att träna och testa taligenkänningssystem
    - alla 11 officiella EU-språk samt samt varianter som finlandssvenska, schweizertyska, walesiska
    - totalt över 60 000 talare inspelade
  - balansera talare enligt
    - dialekt, ålder och kön
  - ca 50 yttranden per talare
    - siffror, datum, tider, penningbelopp, enkla kommandon, fonetiskt rika meningar och ord
  - SpeechDat i Sverige
    - 5000 talare inspelade över vanlig telefon
    - 1000 talare inspelade över mobiltelefon
- GSLT Tal- och taligenkänning M Blomberg 2002-9-9 | 16 |



- Störningar och annat**
- Mobiltelefoni
    - bil, trottoar, restaurang
    - *Bengt Dennis ger inga avskedsintervjuer inför sin avgång vid årsskiftet*
    - *Det handlar bara om ett glapp på 18 månader*
  - Dialektalt uttrycksätt
    - *Han försökte förgäves rädda sin hustru på övervakningen*
  - Den mänskliga faktorn
    - *Kvinnan är mycket nära en total kollaps och gråter upphörligt*
- GSLT Tal- och taligenkänning M Blomberg 2002-9-9 | 18 |

## TMH:s textdatabas

- Totalt ca 150 miljoner ord
- Texter
  - Pressens Bild ca 90 miljoner ord
  - Samhall ca 37 miljoner ord
  - Datalogistik Göteborg ca 20 miljoner ord
  - Göteborgs-Posten ca 5 miljoner ord
- 1,9 miljoner olika ord
- ca 1 miljon ord förekommer bara en gång

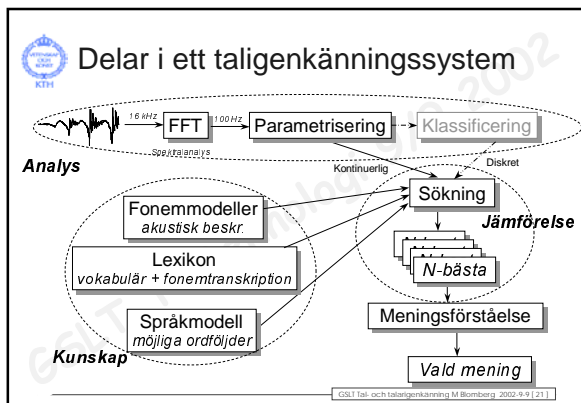
GSL Tal- och taligenkänning M Blomberg 2002-9-9 [19]

## TMH:s textdatabas - de 105 vanligaste orden

Totalt gemensamt		Totalt gemensamt		Totalt gemensamt	
1	4428106 4207712 och	36	341283 319097 eller	71	199430 190186 vill
2	4261937 39023021	37	328109 272105 efter	72	198210 195785 ska
3	3415189 3350874 att	38	321051 310172 ska	73	192338 191218 dam
4	2902952 2048225 det	39	317568 31876 ut	74	191771 190056 blev
5	2643912 2318292 en	40	315375 308648 mot	75	189406 189021 något
6	2530717 2483293 som	41	313483 280863 vid	76	186990 183921 måste
7	2394192 2284315 på	42	310692 281752 har	77	186270 185545 sina
8	2014650 1979468 år	43	307357 299019 också	78	186203 176153 utan
9	1766827 1637863 av	44	306517 249002 du	79	181863 178210 gör
10	1747111 1637525 för	45	303055 299524 är	80	180154 181612 detta
11	1702169 1645599 med	46	302895 258281 under	81	178068 149498 allt
12	1602547 1336856 den	47	298052 221993 då	82	173932 171450 kunde
13	1537177 1493270 så	48	289405 287853 säger	83	172629 164428 kom
14	1412872 1363722 inte	49	289217 283823 över	84	168233 143606 många
15	1396923 997461 han	50	288971 288766 bara	85	166144 153667 någon
16	1354183 1093656 de	51	287357 285420 upp	86	161459 155325 mer
17	1254033 1220682 när	52	278734 241908 alla	87	157873 156289 så
18	1248597 823531 jag	53	274384 199014 ned	88	157639 156759 bit
19	1196578 1099775 om	54	265398 264386 mig	89	154169 152887 till
20	1068768 1041068 var	55	258010 257276 vara	90	154085 144169 första
21	1042099 966735 ett	56	256737 247662 mycket	91	153719 136447 några
22	862144 437264 men	57	253599 250995 in	92	148847 145876 varit
23	785062 783009 sig	58	249388 244958 är	93	144676 141776 fram
24	686040 497508 hon	59	248570 190884 hans	94	144363 131636 hela
25	678986 592543 så	60	244544 232950 andra	95	143998 143564 henne
26	613418 430883 vi	61	243864 237294 för	96	142353 135181 ta
27	573911 544501 från	62	243801 242124 ha	97	141377 130009 genom
28	556489 492627 man	63	243389 193932 sedan	98	140298 138623 mellan
29	550625 541469 hade	64	236593 231545 kommer	99	139927 138478 dag
30	504248 488295 kan	65	224111 221150 ha	100	139762 141499 ingen
31	500352 364182 när	66	222301 221351 honom	101	137551 134286 kronor
32	399292 286295 nu	67	219783 166421 två	102	137291 126642 nya
33	393950 385144 skulle	68	217583 169797 hur	103	135560 135242 göra
34	358713 312987 där	69	207407 202139 finns	104	135335 105644 även
35	348211 344450 sin	70	202274 198088 till	105	134059 301 sveige

ca 150 miljoner ord totalt  
1,88 miljoner olika ord  
1 miljon ord förekommer bara en gång

GSL Tal- och taligenkänning M Blomberg 2002-9-9 [20]



## Olika taligenkänningsmetoder

- **Mönsterigenkänning (Pattern Recognition) (Äldst)**
  - Enkel jämförelse av två spektrala tidsserier
  - Kompensation för varierande talhastighet (Dynamisk programmering, DP)
- **Expertsystem (Övergivet i sin direkta form)**
  - Fonetikers kunskap uttryckt i regler för fonetisk klassning
  - Svårt och inflexibelt
- **Artificiella Neurala Nät (ANN) (Bra för klassificering)**
  - Huvudsakligen för fonetisk klassning
  - Används i hybridssystem tillsammans med HMM
- **Hidden Markov Models (HMM) (Mest använd)**
  - Representerar talets segmentella struktur
  - Viterbi-avkodning (form av DP)

GSL Tal- och taligenkänning M Blomberg 2002-9-9 [22]

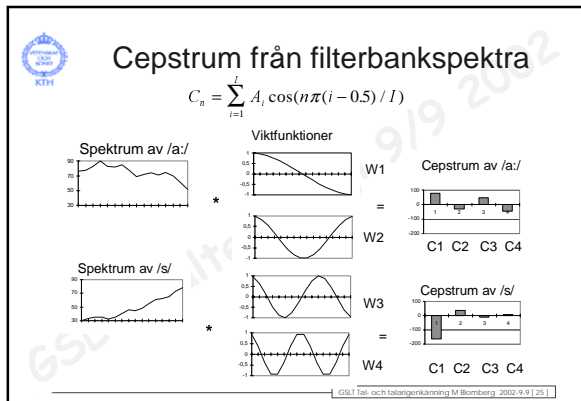
## Parametrar för igenkänning

- Filterbanksamplituder (från FFT, Fast Fourier Transform)
  - Mel-skala - baserad på örats frekvensupplösning
- Cepstrum
  - inversfouriertransform av logaritmiskt spektrum - ortogonala
  - Cepstrum på Mel-spektrum: MFCC - standardmetoden
- LPC
  - linjär prediktion - Linear Predictor Coefficients
- Formanter
  - i kunskapsbaserade system
  - svårt att mäta - kompromiss: mät tyngdpunkter i frekvensband
- Artikulatoriska parametrar
  - nära kopplad till talproduktionen
  - komplicerade att beräkna
- Hörsele baserade parametrar
  - enkel modellering av hörseln
  - förbättring för tal stort av buller och brus

GSL Tal- och taligenkänning M Blomberg 2002-9-9 [23]

## Vanligast: MFCC-analys Mel Frequency Cepstral Coefficients

GSL Tal- och taligenkänning M Blomberg 2002-9-9 [24]

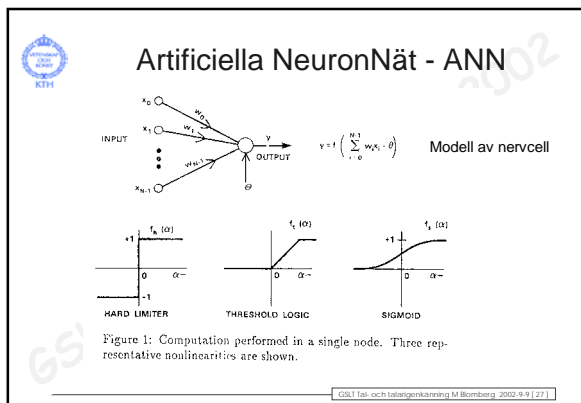


### Vektorkvantisering (VQ)

- Transformering från kontinuerliga till diskreta parametrar
- Automatisk indelning av parameter-rymden i ett litet antal (~256) områden. Minimera distorsion i träningsdata
- Klassa varje tidpunkt av ett yttrande till ett av dessa områden.
- Hela yttrandet beskrivs som en följd av indextal.
  - Kraftig datareduktion på bekostnad av kvantiseringsdistorsion.

Ex. enl trajektorien ovan: 2,1,1,5,5,6,6,6,7

GS&I Tal- och talstämning M Blomberg 2002-9-9 | 26 |



### Artificiella neuronnät - exempel

#### Klassificering av fonemkategorier

- Utlager**  
Aktiveringsgrad för varje kategori
- Dolt lager**
- Inlager**  
Filteramplituder

GS&I Tal- och talstämning M Blomberg 2002-9-9 | 28 |

