# Taligenkänningsmetoder

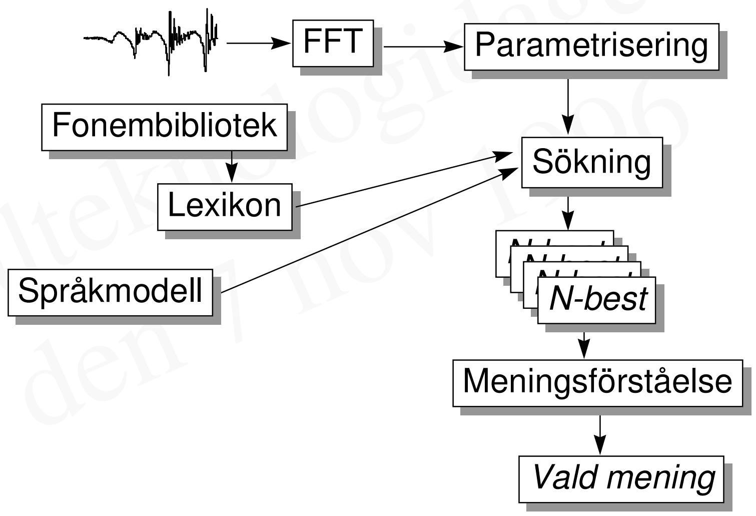
*Mats Blomberg*

*Tal, musik och hörsel*
*KTH, Stockholm*





# Ett generiskt taligenkänningssystem



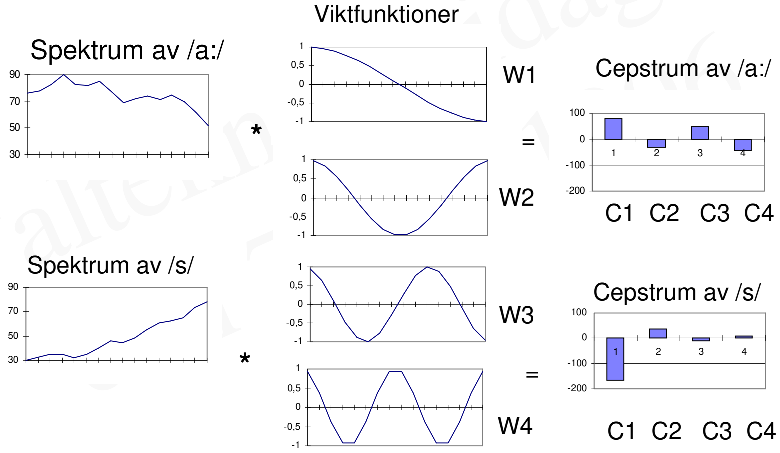
# Vanliga talparametrar

- Cepstrum
  - 12 cepstrumkoefficienter
  - 12 Δ-cepstrumkoefficienter
  - 12 ΔΔ-cepstrumkoefficienter
  - normalisering
    - » subtrahera medelcepstrum över yttrandet
- Energi
  - energi
  - Δ-energi
  - ΔΔ-energi
- Sammanlagt 36 + 3 = 39 inputparametrar för varje spektral mätning var 10:e ms



# Cepstrum från filterbankspektra

$$C_n = \sum_{i=1}^{I} A_i \cos(n\pi(i-0.5)/I)$$



# Representationsformer av fonetisk kunskap

- Spektral tidsserie för varje enhet
- Artificiella neurala nät
  - fonetisk klassificering av spektra
- Markovmodeller
  - sannolikhetsbaserad beskrivning
- Talproduktionsinriktad representation
  - artikulatoriskt inriktad beskrivningsmodell
  - automatisk eller manuell generering





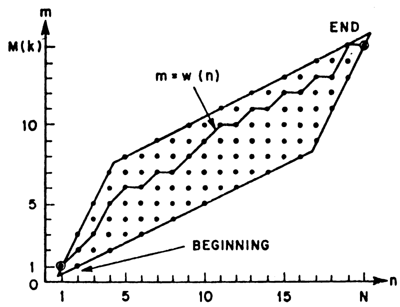
# Klasser av taligenkänningsmetoder

- Mönsterigenkänning (Pattern Recognition)
  - Jämför två spektrala tidsserier
  - Kompensation för varierande talhastighet (Dynamic Time Warping)
- Expertsystem
  - Fonetikerns kunskap uttryckt i regler för fonetisk klassning
  - Svårt och inflexibelt
- Hidden Markov Models (HMMs)
  - Representerar talets segmentella struktur
  - Viterbiavkodning
- Artificiella Neurala Nät (ANNs)
  - Huvudsakligen för fonetisk klassning
  - Hybridsystem med HMM

# Tidsnormalisering vid mönsterigenkänning

Tid

Referens



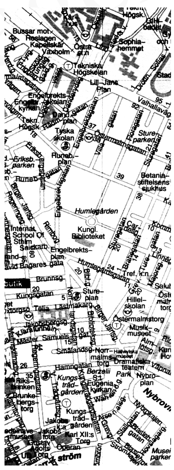

Tid

Okänt yttrande



# Dynamisk programmering

- **Analogi**
  - Sök bästa väg mellan två adresser i en stad.
  - Ett mycket stort antal möjliga vägar, men: vid varje vägval är den bästa vägen till målet oberoende av hur man kom till denna punkt.
  - Den globala optimeringen består av en kombination av oberoende lokala val som kan optimeras var för sig. Antalet sökningar minskas dramatiskt.



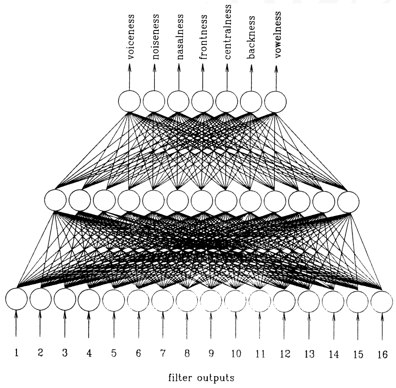
# ANN-exempel

Klassificering av fonemkategorier

**Output layer**
Feature activations

**Hidden layer**

**Input layer**
Filterbank amplitudes






# Hybridsystem - ANN/HMM

- ANN för fonemklassificering, HMM för ord- och meningsigenkänning.
- De bästa hybridsystemen är jämförbara med HMM-system, men med betydligt mer kompakt beskrivning av referensdata.
- Kan vara svåra att träna.

KTH

# Produktionsinriktad representation

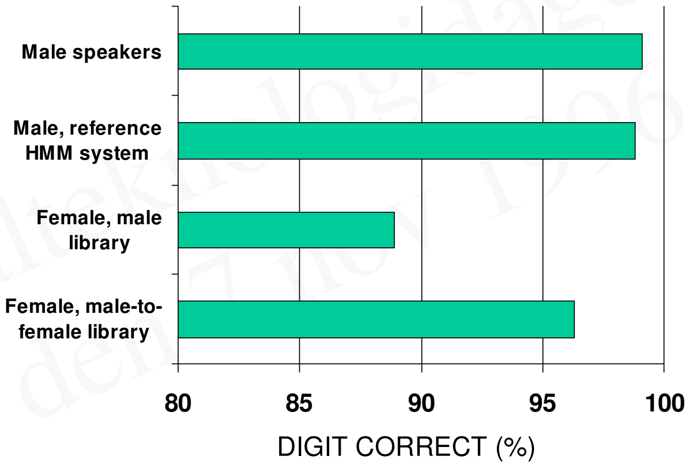
- Fördelar
  - Separat beskrivning av produktionsfaktorer
    - Kan syntetiseras i ett stort antal kombinationer av röstkvalitet, anatomi, talstil, mm.
  - Akustiska och fysikaliska samband begränsar variationen
    - Mindre träningsdata behövs
    - Kan predicera fonemsekvenser som ej tränats
  - “Transformation” av talare
- Problem
  - Parameterextraktion





KTH

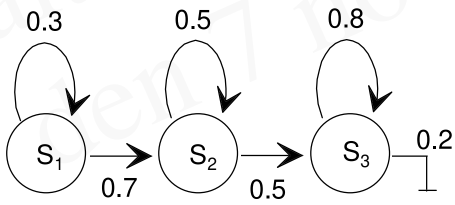
# Transformering från manliga till kvinnliga fonemreferenser med en talproduktionsmodell




(Blomberg, M., ICSLP-94, 1994)

# En Markovprocess

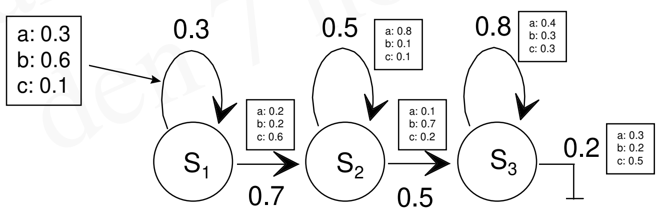
- Modelleras med tillstånd förbundna med bågar med övergångssannolikheter
- Saknar minne
- Förloppet beror enbart på det aktuella tillståndet, ej på historien





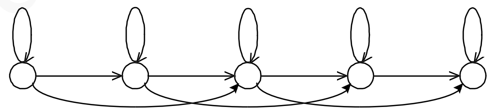

# Hidden Markov Models

- Övergångssannolikheter
- Observationssannolikheter eftersom tillstånden INTE kan observeras direkt
  – *Sannolikheten* för att modellen ska ha genererat en observerad symbolföljd kan beräknas.
  – Den mest sannolika tillståndsföljden (Viterbi) eller de N mest sannolika följderna (N-best algorithm) kan bestämmas



# HMM för taligenkänning

- Talsignalen betraktas som genererad av en HMM och beskrivs av
  - **Övergångssannolikheter**
    - Resulterar i en enkel modell av segmentdurationer: exponentialfördelning
  - **Observationssannolikheter**
    - Statistiska modeller av talljudens akustik
- Snabbare är DTW pga sin segmentella, fonetiska struktur
- Lämplig för stora vokabulärer och kontinuerligt tal





# Representation av fonemmodeller

- Kontextberoende allofoner
  - trifoner, difoner, monofoner beroende på om båda, ena eller inget av de omgivande allofonerna är bestämd
- Varje allofon är ofta uppdelad i flera tillstånd som beskriver t. ex. initiala, mediala och finala delar.
- Varje tillstånd bestäms med fördelningar för observations- och övergångssannolikheter

# Varianter av HMM för igenkänning

- Diskret HMM
- Kontinuerlig
- Halvkontinuerlig (Semi-continuous)
- Semi-Markov (bättre durationsmodell)
- Enkel/sammansatt sannolikhetsfördelning (mixture density)
- Delning och knytning
  - Olika modeller har gemensamma delar





# Diskret HMM

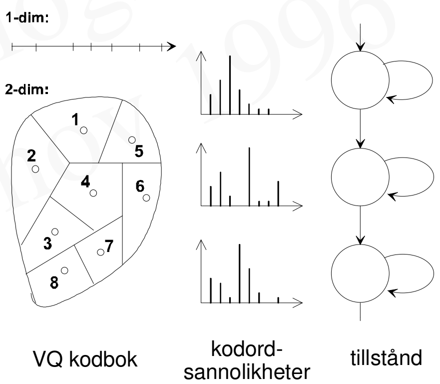
- Begränsat antal observationssymboler med diskreta sannolikheter
- Talet beskrivs med vektorkvantisering (VQ)

+ godtycklig statistisk modell

- kvantiseringsdistorsion

- VQ och HMM optimeras separat

- kräver mycket träningsdata



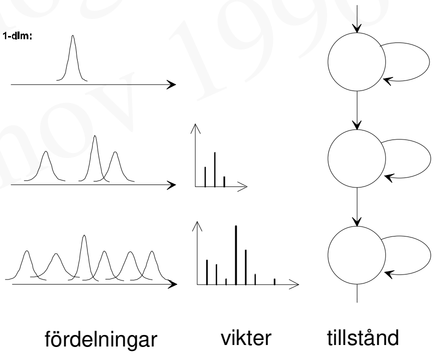
KTH

# Kontinuerlig HMM

- Kontinuerliga observationsvariabler (akustiska mått)
- En eller flera sammanvägda Gaussfördelningar (mixture densities)

+ bättre modell av observationernas fördelning
+ kräver mindre mängd träning vid få komponenter

- Flera Gausskomponenter krävs normalt
- beräkningsintensiv





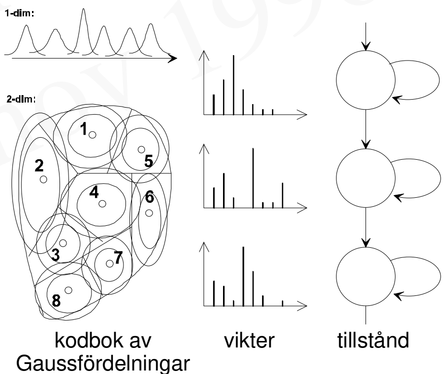

KTH

# Halvkontinuerlig HMM

## Semi-Continuous HMM, SCHMM

- Kontinuerliga observationer
- Alla tillstånd delar en kodbok av Gausskomponenter med individuella vikter

+ samtidig optimering av VQ och HMM
+ antalet fria parametrar *kan* minskas
+ beräkningsmängden *kan* minskas



# Sökmetoder med HMM

- Viterbi-algoritm
  - beräknar sannolikheten för den bästa tillståndsföljden med *dynamisk programmering*.
  - Används ofta.
- Forward-algoritm
  - beräknar sannolikheten att en given HMM producerat indata
- Forward-Backward-algoritm
  - används vid träning för att beräkna maximum likelihood uppskattning av HMM-parametrar med "reestimering" (en iterativ metod).





# Knytning (tying) av HMM-modeller

- Träningsdata är vanligtvis ojämnt fördelad
  - Svårt att träna modeller som har få observationer
  - Exempel: trifoner
- Lösning: Vissa tillstånd eller Gausskomponenter kan vara gemensamma för flera modeller.
  - Exempel: trifonerna t-a-m och d-a-m kan dela finala tillstånd.
  - \+ Antal parametrar minskar
  - \+ Mer träningsdata per parameter
  - \+ Minnesstorleken minskas
- Knytningar kan göras manuellt eller med automatiska metoder

# Miljoner parametrar

- System med stor vokabulär och ordgränstrifoner använder ~ 60 000 trifoner
- Akustisk vektor med 39 element och 10 Gauss-komponenter ger 790 parametrar per tillstånd.
  - 39 x 2 x 10 + 10 = 790
- 60 000 trifoner med 3 tillstånd ger 142 000 000 parametrar!
- Om knytning (*tying)* används för att dela modell-parametrar.
  - ger ~10 000 000 parameters.



# Kombinera akustisk och språklig kunskap: Bayes regel

$$P(Words\,/\,Acoustics) = \frac{P(Acoustics\,/\,Words) \times P(Words)}{P(Acoustics)}$$

- *P(Words/Acoustics)* är *a posteriori sannolikheten* för en ordföljd givet den akustiska informationen.
- *P(Acoustics/Words)* är *sannolikheten* att ordföljden genererar den akustiska informationen och beräknas i ett träningsmaterial.
- *P(Words)* ges av språkmodellen och är *a priori sannolikheten* för ordföljden (N-grams).
- *P(Acoustics)* kan ses som en *konstant* eftersom den är oberoende av ordföljden och kan ignoreras

KTH

# Språkmodeller för igenkänning

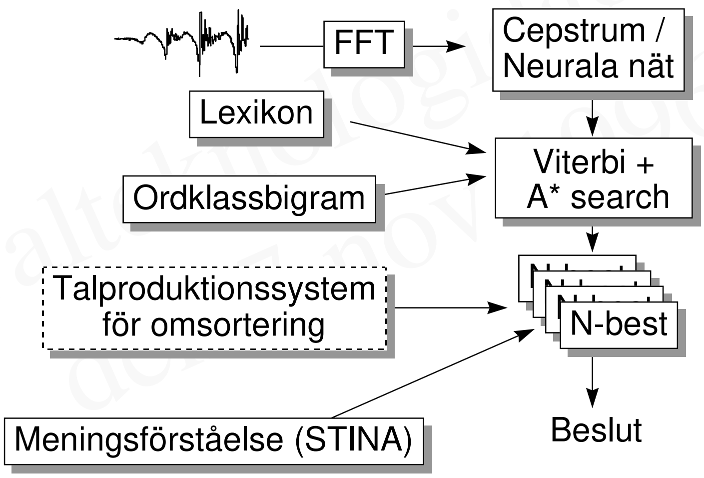


- N-gram (ordföljdssannolikheter) ger bra resultat trots sin enkelhet
  - bigram: $P(w_i \mid w_{i-1})$
  - trigram: $P(w_i \mid w_{i-2}, w_{i-1})$



KTH

# WAXHOLM-systemet




Talteknologidagen den 7 nov 1996