Forskningsläge internationellt, marknad och tillämpningar

Kjell Elenius
Institutionen för tal, musik och hörsel
KTH, Stockholm

SpeechDat

- Samla in tal över telefonnätet för att kunna träna taligenkänningssystem
- Alla EU-språk samt några till
  - 15 olika språk med 21 varianter
    - svenska, finlandssvenska
    - danska, norska, finska
    - tyska, schweizertyska, luxemburgtyska
    - holländska, flamländska
    - brittisk engelska, walesiska
    - franska, belgisk franska, schweizerfranska, luxemburgfranska
    - spanska, italienska, portugisiska
    - grekiska, slovenska
- Finansieras av EU:s 4:e ramprogram
  - TMH är "associated partner" till Vocalis i England och får 128 000 ECU under 2 år; 1996-98

KTH

SpeechDat

- Ungefär 40 yttranden spelas in i ett samtal som varar högst 10 minuter per talare
  - enstaka siffror och sifferkombinationer
  - kommandoord och fonetiskt rika ord
  - 9 fonetiskt rika meningar
  - några spontana fraser
  - person- och företagsnamn
- Talarna fördelas efter
  - 2 kön, 4 ålderskategorier, 18 dialektområden för hela Sverige
- 3 delprojekt
  - vanliga telefonnätet, mobiltelefonnätet, talarverifiering
- TMH spelar in
  - 5000 talare över vanliga telefonnätet
  - 1000 talare över mobiltelefonnätet
    - 900 GSM och 100 NMT

KTH

TMH:s textdatabas

- Totalt ca 150 miljoner ord
- Texter
  - Pressens Bild ca 90 miljoner ord
  - Samhall ca 37 miljoner ord
  - Datalingvistik Göteborg ca 20 miljoner ord
  - Göteborgs-Posten ca 5 miljoner ord
- 1,88 miljoner olika ord
- ca 1 miljon ord förekommer bara en gång

KTH

Teknik för taligenkänning

- HMM - Hidden Markov Models
  - Dominerande metod, baserade på inlärning, men inbegriper även fonetisk kunskap
    - Cambridge Univ. (Steve Young)
    - LIMSI, Paris (Jean-Luc Gauvin)
    - SRI, Stanford
    - CMU, Pittsburgh
- ANN - Artificial Neural Networks
  - inte lika populära som för något år sedan
  - Hybridsystem, ANN & HMM , i klass med bästa HMM-systemen
    - Cambridge Univ. (Tony Robinson)
    - ICSL (Bourlard, Morgan)
- Kunskapsbaserade metoder
  - kombineras med automatisk inlärning
    - MIT
    - KTH

KTH

Aktuella forskningsområden för taligenkänning

- Dialogsystem, några exempel
  - MIT, Galaxy: talstyrd webbläsare
  - SRI & Telia Research: flygbokning (ATIS-domänen)
  - LIMSI, Paris: Multimodal-Multimedia Automated Service Kiosk (MASK) - tåginformation och biljetter
  - Philips: upplysning om tågtider, i funktion i Schweiz
  - KTH: Waxholm - skärgårdsinformation
- Verbmobil, Tyskland
  - stöd för människor som talar olika språk: tyska och japanska
  - 150 forskare från 29 olika organisationer

Igenkänning av tal från radio och video, indexering

KTH

- Några artiklar från ICSLP 96 i Philadelphia, oktober 96
  - Gelin & Wellekens, Institut Eurécom, Sophia Antipolis: *Keyword spotting enhancement for video sountrack indexing*
    - nyckelordsigenkänning, talaroberoende
    - känna igen ett godtyckligt ord i en öppen vokabulär
      - “tennisresultat”, “Roxette”, “ “Jeltsin“
    - indexeringen indelas här i två steg
      - först igenkänning som ger möjliga talljud, som sedan används vid själva frågetillfället
  - Spina & Zue, MIT: *Automatic Transcription of General Audio Data: Preliminary Analyses*
    - analyserar bland annat hur vokabulären växer över tiden
  - Kubala, Anastasakos, Jin, Nguyen & Schwartz, BBN: *Transcribing Radio News*
    - försöker känna igen alla ord som sägs på radion

Språkinlärning

KTH

- Några artiklar från ICSLP 96 i Philadelphia, oktober 96
  - Yoram &Hirose, University of Tokyo: *Language training system utilizing speech modification*
    - känna igen feluttal och försöka accentura det korrekta uttalet på samma sätt som en vanlig lärare
      - sänkt talhastighet
      - överdriven intonationskontur
      - ökad intensitet
  - Neumeyer, Franco, Weintraub & Price, SRI: *Automatic text-independent pronunciation scoring of foreign language student speech*
    - anpassat taligenkänningen till att ge ungefär samma bedömning av uttalet som som en människa gör
      - visar att relativa fonemlängder ger en bra uppfattning om uttalet

ARPA-projektet i USA

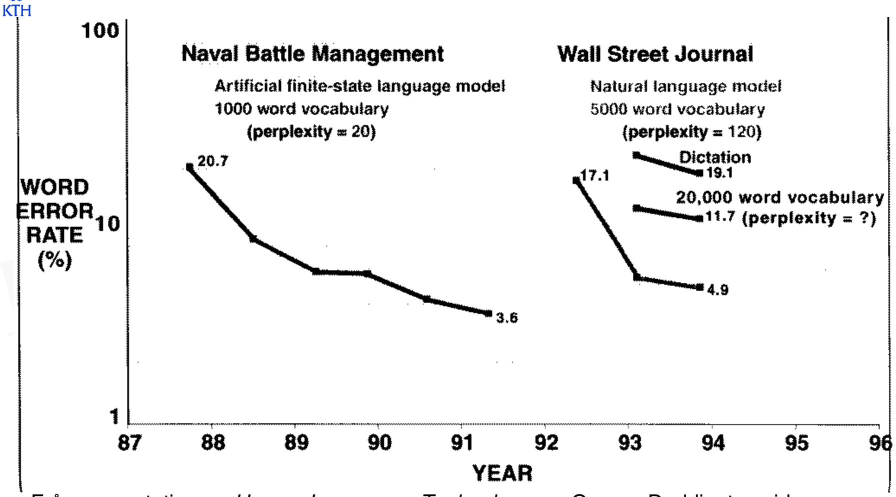
- Advanced Projects Research Agency
  - började 1984
  - deltagare
    - CMU, SRI, BBN, MIT, Lincoln Labs, Dragon Systems
  - "competative evaluations" varje år
  - domäner
    - Naval Battle Management, ~1000 ord
    - ATIS, Air Traffic Information System, ~1000 ord
      - Flygbokning
    - Wall Street Journal, 5000 - 60000 ord
      - Tidningstext, uppläst
    - SWITCHBOARD
      - samtal över telefon med okänd person om givet ämne
    - CALL HOME
      - samtal över telefon med närmaste familjekretsen

svårare

Switchboard-databasen

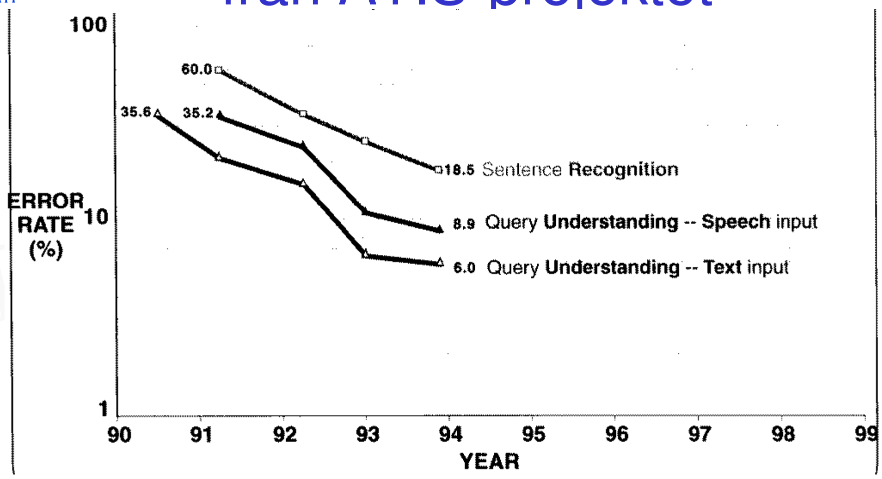
- "There is no data like more data"
  - insamlat av Texas Instruments under tidigt 90-tal
  - "spontant" tal över telefon mellan för varandra okända personer
  - 70 olika samtalsämnen
  - ursprungligen för talarverifiering och "ämnesidentifiering"
    - sedan även för taligenkänning
  - omkring 2500 samtal, 3 - 10 minuter långa
    - 500 talare från hela USA
      - över 1 timmes tal från 50 “måltalare”
  - 150 timmar tal
    - 2 000 000 ord
  - naturlighet 1,48; 1 mycket naturligt - 5 “ansträngt, artificiellt”
  - målsättning: detaljerad, fonetisk uppmärkning av allt tal

ARPA-resultat



Från presentationen, *Human Languages Technology*, av George Doddington vid ARPA Workshop, New Jersey, 8 - 11 mars, 1994.

ARPA-resultat från ATIS-projektet



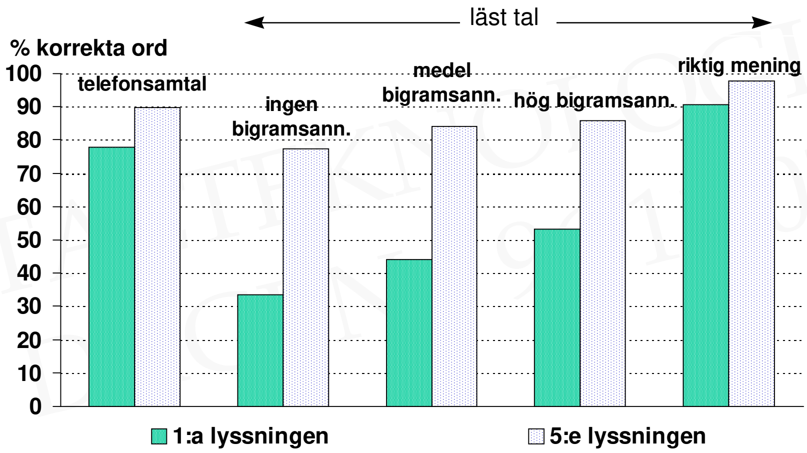
Från presentationen, *Human Languages Technology*, av George Doddington vid ARPA Workshop, New Jersey, 8 - 11 mars, 1994.

Igenkänningsresultat för olika ARPA-databaser

Telefonresultaten kommer från John Makhouls föredrag vid IVTTA 96, Interactive Voice Technology for Telecommunications Applications, Basking Ridge, New Jersey, 30/9 - 1/10 -96.

Databas	Inspelade timmar (för inlärning)	Antal ord för inlärning av språkmodellen	Ordfel
Wall Street Jour. kontor telefon	150 (60)	200 000 000	 7% 37%
Switchboard telefon	150 (70)	2 000 000	 38%
Call Home telefon	13	200 000	 50%

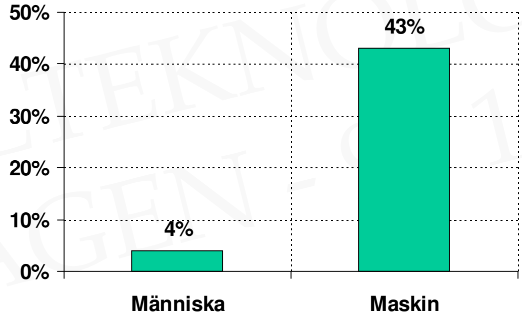
Mänsklig förståelse av Switchboard-meningar



Cole, Yan & Bailey: The influence of bigram constraints on word recognition by humans: implications for computer speech recognition, ICSLP 96, Philadelphia, Oct. 1996.

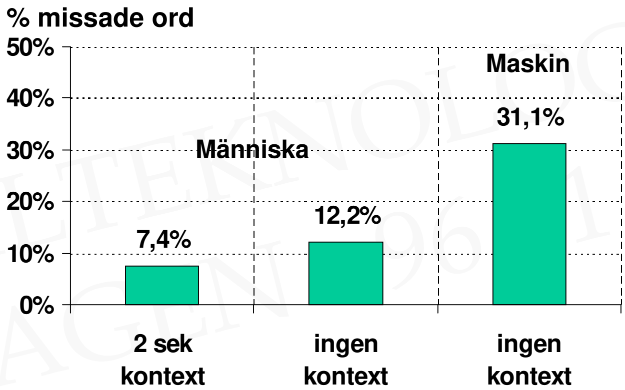
Kreditkortsämnet i Switchboard

Resultat för mänsklig transkription av hela samtal jämfört med automatisk igenkänning av extraherade meningar från samma material. Notera att detta maskinresultat är något tidigare och därför något sämre än vad som visas i bild 13.



R. P. Lippman: "Speech perception by humans and machines", Proc. of the workshop on the auditory basis of speech perception, Keele University, UK, 15-19 July, 1996, pp. 309-316.

Kreditkortsämnet i Switchboard hitta 20 nyckelord



R. P. Lippman: "Speech perception by humans and machines", Proc. of the workshop on the auditory basis of speech perception, Keele University, UK, 15-19 July, 1996, pp. 309-316.

Memento

- Ordfelprocent för ett system från AT&T
  - Vuxna 1%
  - Barn 30%

Några taligenkänningssystem för persondatorer

- Styrning av vanliga applikationsprogram samt diktering med vokabulär om 30 - 60 000 ord.
  - IBM
    - amerikansk och brittisk engelska samt flera europeiska språk
  - DragonDictate
    - amerikansk och brittisk engelska samt flera europeiska språk, bland annat svenska
  - Kurzweil
    - amerikansk engelska
  - Apple
    - kinesiska
  - Microsoft
    - fortfarande under utveckling

Taligenkänning i telefonbolag -några exempel från USA

- AT&T och Lucent Technologies
  - VRCP, "Voice Recognition Call Processing", val av betalningssätt
    - vokabulär om 5 ord: collect, calling card, third number, person, operator
    - kompletteras nu med sifferigenkänning av telefon- och kreditkortsnummer
    - mer än 5 miljoner samtal per dag
  - AVL, AT&T Voice Line
    - försök med röststyrd uppringning, säker kontoaccess m.h.a. kontonummer och talarverifiering, röststyrd administration av teleabonnemang, m.m.
    - testat med 400 personer
  - Universal Card Services
    - igenkänning av enstaka siffror, påminner om telefonbank med knappsats
    - 1000-tals användare per dag
- Nynex
  - VoiceDialing Service
    - 18 talaroberoende ord, 75 talarberoende namn

Taligenkänningsföretag i USA - huvudsakligen för telefoni

- Exempel på några mer renodlade taligenkänningsföretag
  - ALTech (MIT)
    - demo: beställning av CD-skivor
  - Entropics
    - säljer HTK (Hidden Markov Toolkit, från Cambridge Univ., UK)
  - Nuance (SRI)
    - aktiekurser, 13000 namn
  - PureSpeech (CMU, MIT)
    - Juggler - personlig assistent
      - för Pentium Pro, inkluderar e-post, fax, röstbrevlåda
  - Texas Instruments
  - Voice Control Systems (VCS)
  - Wildfire Communications, Inc.
    - Intelligent Electronic Assistant

Personlig telefonassistent

- Marknadsförs bland annat av ALTech, PureSpeech och Wildfire
  - Exempel på meningar man kan säga till Juggler från PureSpeech:
    - "Call my sister on her cellular phone."
    - "Describe my faxes."
    - "Please read my e-mail."
    - "Find a name in my address book."
    - "Leave a message for Susan."

Taligenkänning i telefonbolag -några exempel från Europa

- England
  - BT: CallMinder
    - byggs av Ericsson, "telefonsvarare i nätet", 500 000 användare
- Frankrike
  - CNET: talstyrd röstbrevlåda
    - vokabulär 7 ord: ja, nej, lyssna, spela in, radera, addera & meddela
    - finns nu i 3 städer, 240 linjer, men under 97 ska hela landet anslutas
- Italien
  - CSELT: nummer-till-namnupplysning
    - 50 miljoner samtal per år, omkring 10% beräknas använda taligenkänning som installeras under 1996
- Tyskland
  - Deutsche Telekom: FAUST - försök med nummerupplysning
    - nyckelordsigenkänning, bokstavering
    - 5000 försökspersoner

Exempel på företag i Europa som arbetar med taligenkänning

- Belgien
  - Lernout & Houspie
    - enstaka ord, högst 200, bruståligt
      - för bilar och konsumentprodukter
    - kontinuerligt tal, 1000 ord, med ekosläckning
      - för datorer, multimedia och telefoni
- England
  - Vocalis
    - *Operetta*™: "automatisk telefonist"
- Frankrike
  - Vecsys
    - *Parole*: simulerar "flygande piloter" vid flygledarträning
- Tyskland
  - Daimler-Benz AG
    - röststyrd telefon kan levereras i S-klass-bilar från nov -96

IEEE Workshop on ASR, Snowbird, Dec. 1995

Taligenkänningsgenombrott under senaste 5 åren

Resultat av enkät bland närvarande talforskare

- Minne/CPU
- Programteknologi
- Mer data
- EM (Expectation Maximization) algoritmen
- Beslutsträd (Context decision trees)
- Söktekniker
- Jämförande utveckling (competitive evaluation)
- Ingenting

KTH

Vad kommer att ge framtida genombrott?

- Fonetisk kontext-modellering
- Fonetiska särdrag
- Bättre tidsupplösning
- Modellera akustiska samband över längre tider
- Prosodi, dynamiska förlopp

- Mer data
- Oövervakad inlärning
- "Faktoriserade modeller"
- Samordnade projekt
  samarbete och infrastruktur
- Dialogsystem
- Paradigmskifte
- Diagnostisk utvärdering
- $$-succé

KTH

Människan och maskinen

Resultat av omröstning: "Om hur många år får maskinen samma resultat som människan för akustisk/fonetisk igenkänning?"

60 talforskare svarade:

Antal år	Antal röster
5 - 10	10
10 - 20	30
20 - 40	10
> 40	10

Några nya idéer

- *Towards increasing speech recognition error rates,* artikel i Speech Communication, vol 18, nr 3, May 1996 av Hervé Bourlard, Hynek Hermansky och Nelson Morgan:
- Propagerar för att man måste kunna testa helt nya metoder och tekniker och att detta måste få ta sin tid även om det inte fungerar lika bra som tidigare använda metoder som naturligtvis är mycket mer optimerade (jämför med ARPA-projektet).
  – Till exempel har de fått relativt goda resultat med att basera igenkänningen bara på transitionerna mellan talljuden medan de helt ignorerar de stationära delar av talsignalen.
  – De gör även försök med att dela in talspektrum i 3 olika oberoende kanaler.

Ett korn av sanning?

- Citat av Fred Jelinek i en kommentar till föregående artikel från samma nummer av Speech Communication:
- *"...electronic technology directly improves only productivity, and often lowers costs by simply shifting them from the provider to the user. Indeed, speech recognition is sold on the basis of its potential to displace operators, but will not enhance their quality of life, while the net effect on the customer may well be negative.*
  I think that very large vocabulary recognizers may have a better chance to at least do no harm. For instance dictation machines, or foreign language teachers."