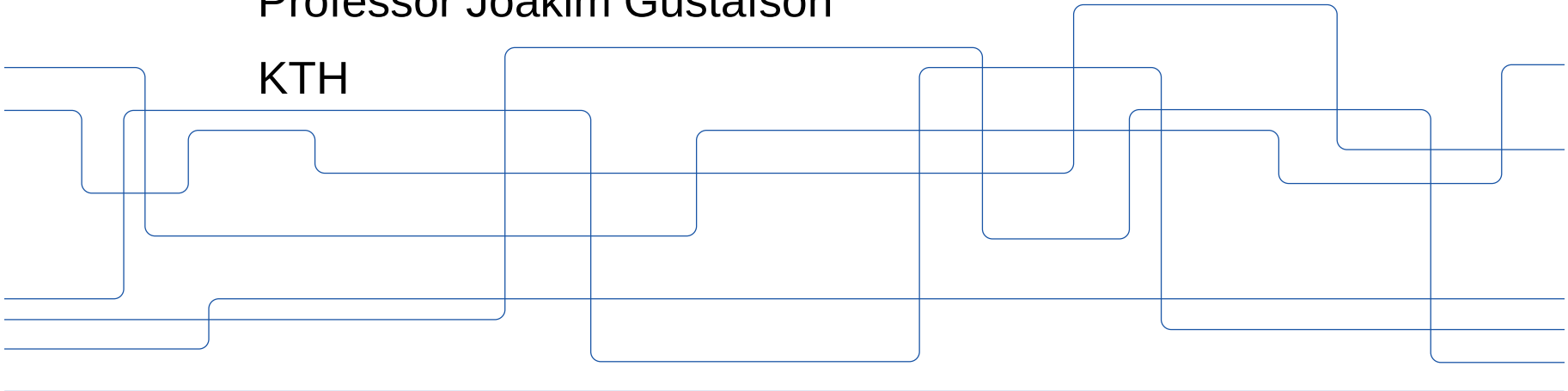


INTELLIGENTA SOCIALA ASSISTENTER

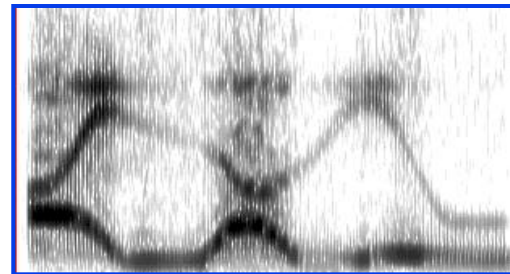
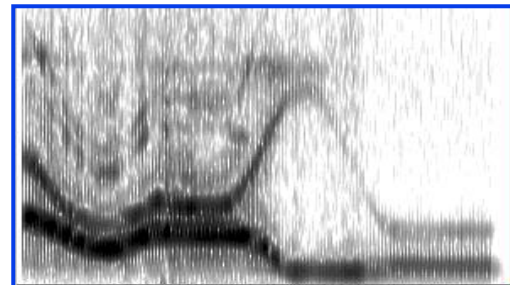
Professor Joakim Gustafson

KTH

A series of thin, light blue lines that form a complex, overlapping pattern across the lower half of the slide. The lines are horizontal and vertical, creating a series of rectangular shapes that vary in size and position, resembling a stylized staircase or a network diagram.

Institutionen för Tal, Musik och Hörsel (TMH)

grundad av Gunnar Fant 1951



Gunnar Fant and OVE I 1953

Ljudexempel: <https://acousticstoday.org/klatts-speech-synthesis-a/>





Anställda på TMH 2022

Professorer



Emereti



Lektorer



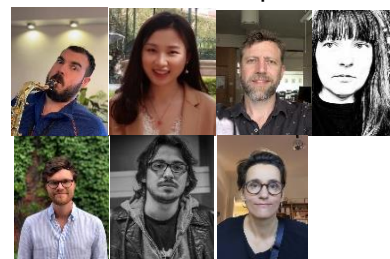
Forskare



Doktorander



Postdocs/Teknisk personal





Vår forskning

förstå hur människor kommunicerar via tal

- Mänsklig perception/produktion av tal
- Multimodal interaktion
- Prosodi i samtal
- Samtalande system
- Datainsamlingar och infrastruktur



Vår vision om samtalande maskiner

Man ska kunna tala med maskiner på samma sätt som man talar med andra människor



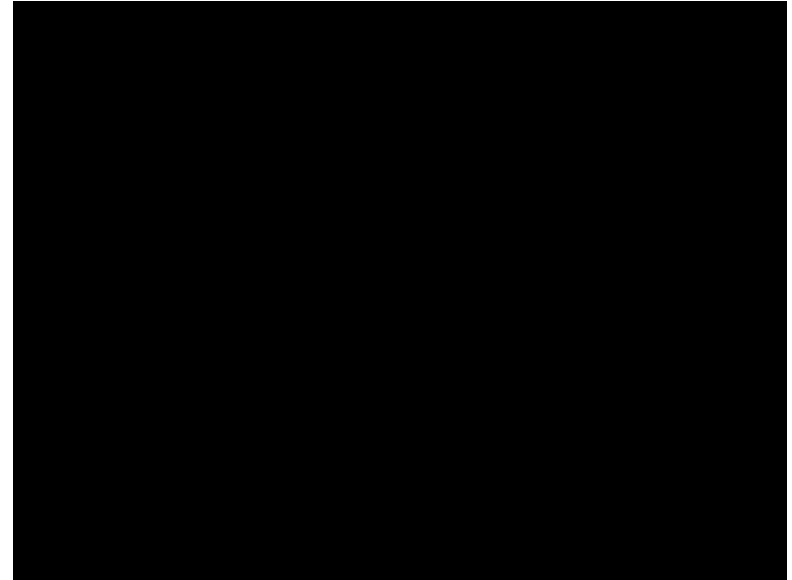
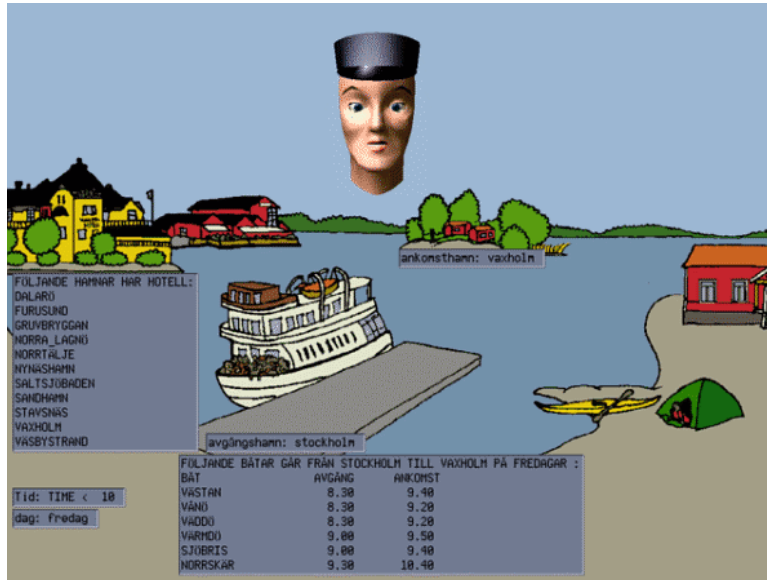
Telia Research 00-07

KTH 07-



Waxholm

det första svenska dialogsystemet 1995



<https://youtu.be/kEuw9zB1u5I>

August

En virtuell August Strindberg på Kulturhuset 1998



https://youtu.be/S_17Fp-wgQk



Furhat

KTHs sociala robot 2010



<https://youtu.be/EzdfvRLUgqY>



Dagens ämne

Intelligenta Sociala Assistenten

Med andra ord:

Maskiner som kan **hjälpa** människor
på ett **smart** och **naturligt** sätt

Artificiell Intelligens

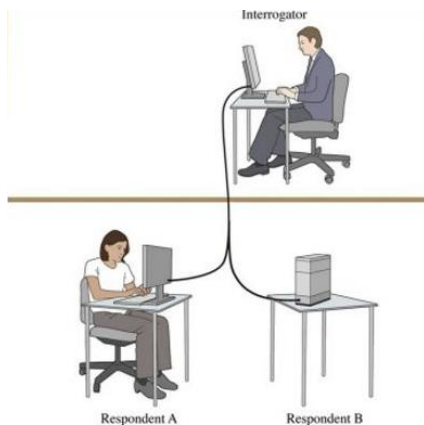
Maskiner som kan efterlikna människors förmåga att förstå språk och som kan lära sig att hitta generella lösningar på problem

- **AI födelse** - vid konferens på Dartmouth College 1956.
- **AI-vinter** - intresset svalnar 1974–1980 och 1987–1993
- **AI-vår** - intresset ökar igen 2010-?

Hur kan vi veta om en maskin lyckats efterlikna mänsklig intelligens?

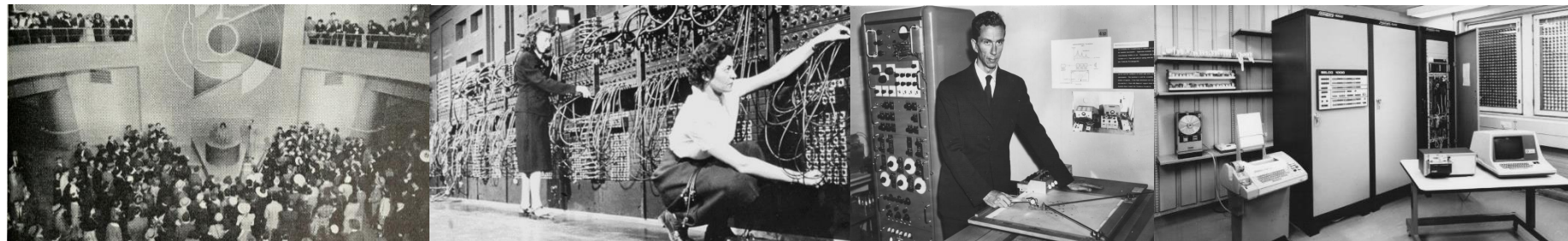
Alan Turing formulerade Turingtestet 1950:

”om en människa konverserar med en maskin och inte kan avgöra om det är en maskin eller människa, uppfylls kriteriet för mänsklig intelligens.”



Praktiskt går det till så att en person sitter vid en dator i ett rum och ställer vilka frågor som helst till andra datorer via nätverk. Dessa är antingen styrda av människor eller datorprogram som svarar på frågorna. Om testpersonen efteråt inte kan avgöra vilka respondenter som är människor och vilka som är program sägs de sistnämnda ha klarat testet.

Vägen mot intelligenta maskiner



Voder (1938)

Första elektroniska talsyntesen

ENIAC (1945)

Första programmerbara datorn

OVE (1953)

Talsyntes från KTH

ELIZA (1964)

Tidig chatbot från MIT



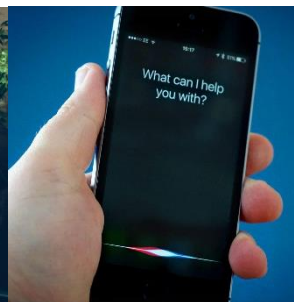
Speak & Spell (1978)

Laptop lär barn att stava



Deep Blue (1997)

slår stormästare i schack



Siri (2011)

Röstagent i mobil



Alexa (2014)

Smart högtalare



AlphaGo (2016)

Slår mästare i Go



Idag talar vi om fyra slags artificiell intelligens

	<i>Ingen människa medverkar</i>	<i>Människa medverkar</i>
<i>Regelstyrt system</i>	Automatisk Intelligens	Assisterad Intelligens
<i>Adaptivt system</i>	Autonom Intelligens	Förstärkt Intelligens



Automatisk Intelligens



*Regelstyrt System
Ingen människa medverkar*

Assisterad Intelligens



*Regelstyr System
Människa medverkar*

Automatisk Intelligens



*Adaptivt System
Ingen människa medverkar*

Förstärkt Intelligens

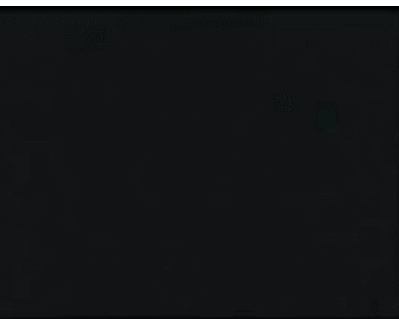


*Adaptivt System
Människa medverkar*

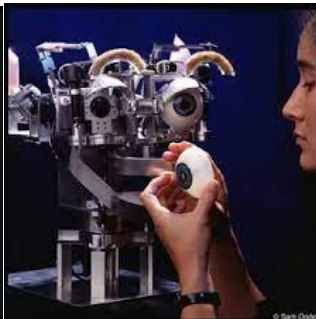
Vägen mot assisterande robotar



Metropolis (1927)
Första filmroboten



Electro (1939)
Världsutställningen New York



Kismet (1998)
Emotionell robot från MIT



ASIMO (2000)
Humanoid robot



PARO (2001)
Robot för dementa



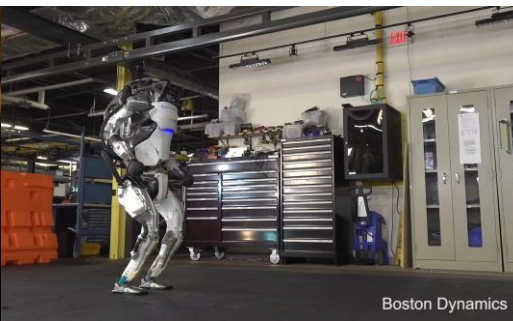
NAO (2008)
Robot för barn



Geminoid (2008)
Fjärrstyrd android



Furhat (2010)
Social robot från KTH



Atlas (2013)
Robot för katastrofer



Pepper (2014)
Robot som guide

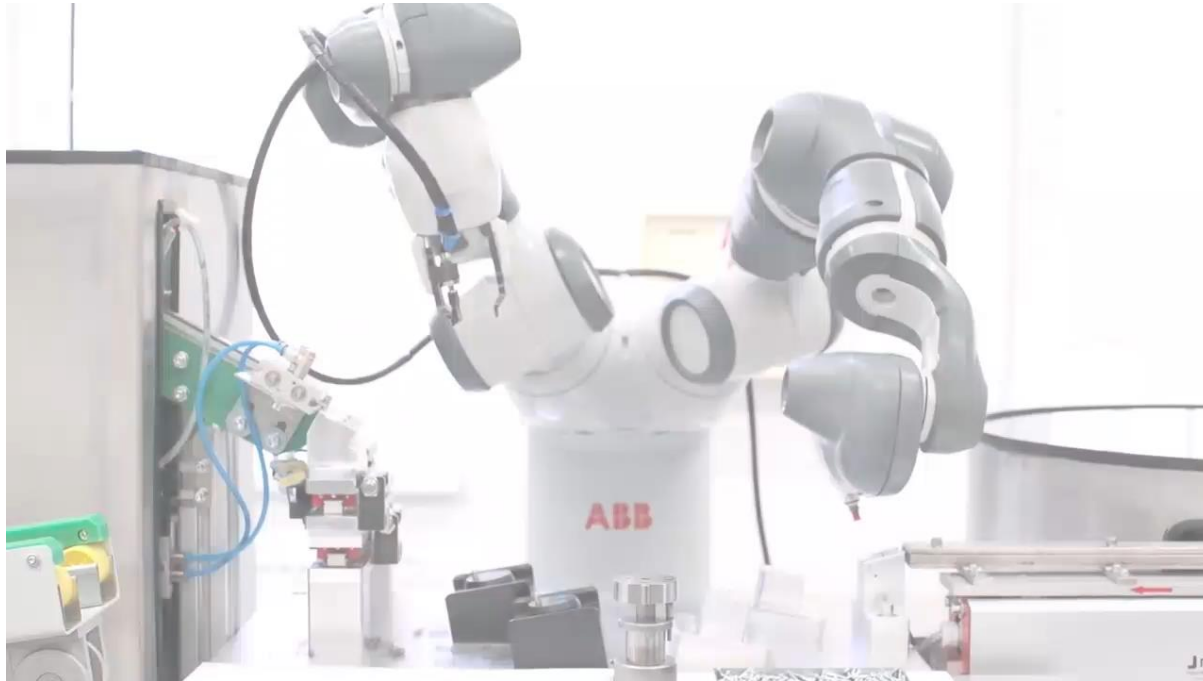


Sophia (2016)
Robot som sällskap

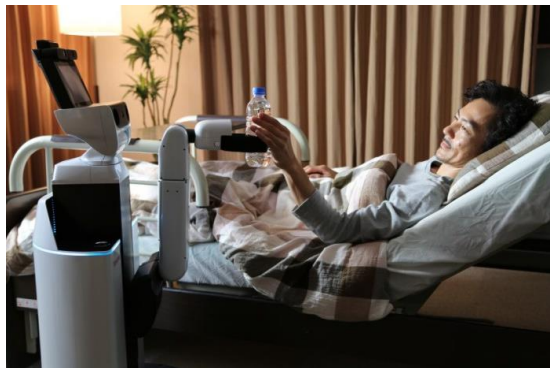


TeslaBot (2023)
Framtiden?

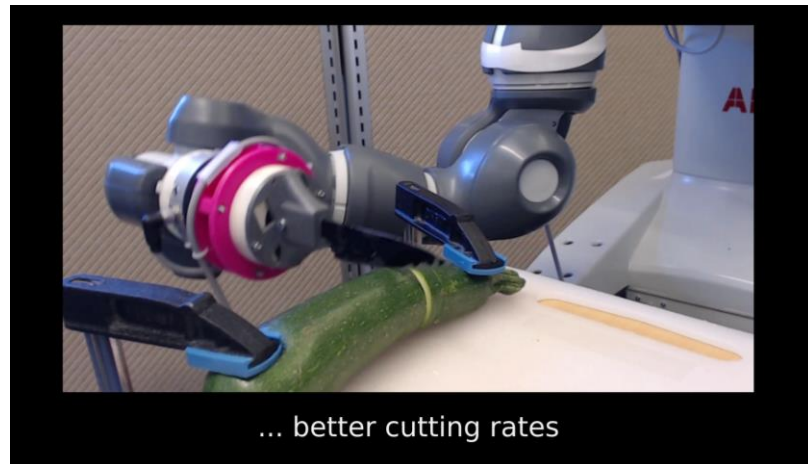
Samarbetande industrirobotar



Assisterande robotar i hälso- och åldringsvård



Hjälpssamma robotar i köket?



Lärarrobotar för autistiska barn



Lek och lär-robotar för barn?



En social robot att plugga med?





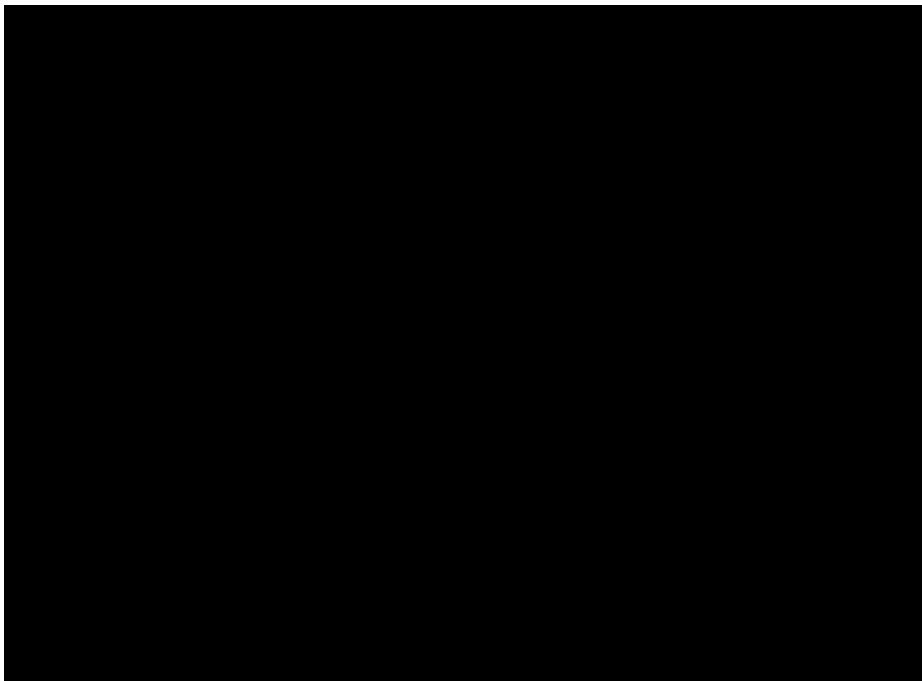
Sociala robotar behöver mänskliga egenskaper

- En kropp som passa mänskliga miljöer
 - ben eller hjul
 - Armar och händer
 - Huvud med ögon och mun
 - Sinnen
 - Syn
 - Hörsel
 - Känsla
 - Smak och luktsinne?
 - Intelligens
 - Förstå och lösa problem
 - Förstå och generera mänskligt tal
-



Vad rösten förmedlar

- **Orden / innehåll**
(Den lingvistiska komponenten)
- **Identitet**
(Den extralingvistiska komponenten)
- **Attityd och emotion**
(Den paralingvistiska komponenten)
- **Val av nästa talare**
(Dialogkontrol)



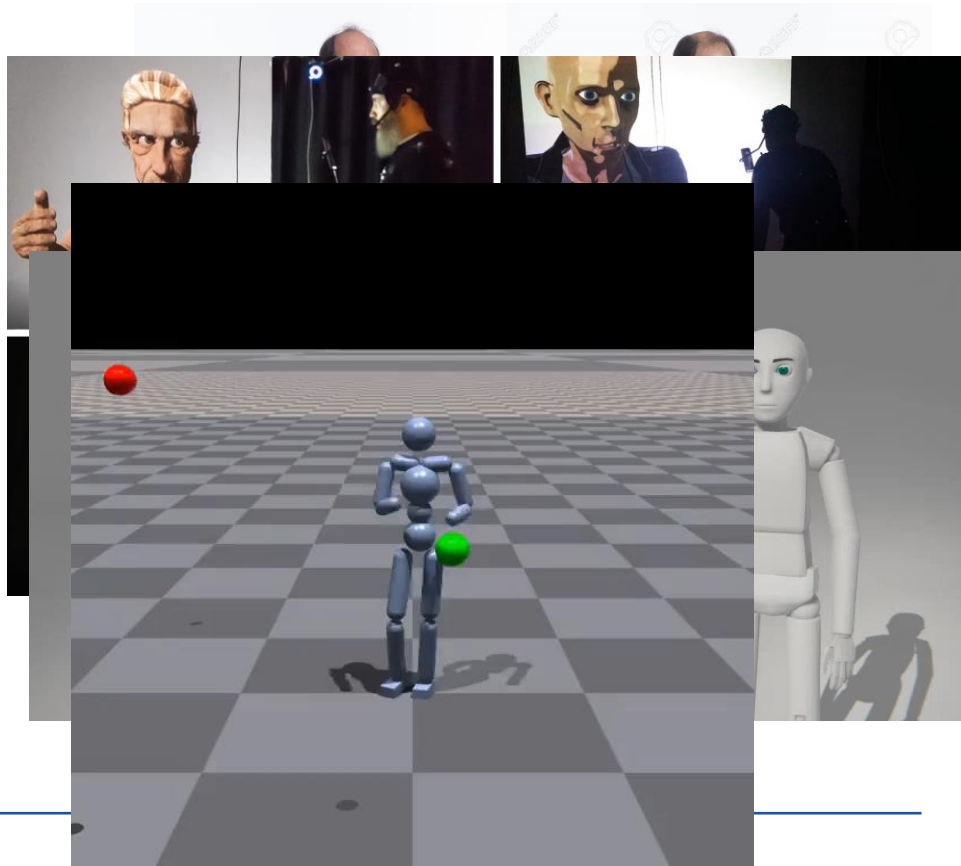
Vad ansiktet förmedlar

- **Orden /innehåll**
(Läpprörelser underlättar förståelse)
- **Identitet**
(Ansiktetsformen m.m. ger identitet)
- **Attityd och emotion**
(Ansiktuttryck förmedlar känslor och åsikter)
- **Val av nästa talare och fokus**
(Blickriktning signalerar uppmärksamhet)



Vad kroppsspråket förmedlar

- **Orden / innehåll**
(Ikoniska och metaforiska gester)
- **Identitet**
(Personliga rörelsemönster)
- **Attityd och emotion**
(Kroppshållning och rörelsemönster)
- **Fokus**
(Pekande signalerar plats)





Det är viktigt att talgränssnitten är intuitiva

- Användaren ska inte behöva lära sig:
 - *När* man kan tala
 - *Hur* man ska tala
 - *Vad* man kan säga
 - Maskinen måste därför:
 - Klart visa när den lyssnar
 - Förstå när användaren har talat klart
 - Säga sånt som speglar det den förstår
-



Samtalskontroll

- Ett samtal har två kanaler
 - Själva informationsutbytet
 - Kontroll av informationutbytet
 - Samtalskontroll
 - Turtagning: avgöra vem som för tillfället har rätt att tala
 - Återkoppling: visa att man lyssnar, att man förstår och vad man tycker om det som sägs
-

Ansiktsgester för samtalskontroll

MonAMI reminder

- **Input:** trådlös mikrofon och webkamera
- **Output:** talande huvud på skärm
- **Samtalskontroll:** lyssnar bara på sin användare när denne tittar på skärmen



Blickbeteende för samarbete

- BabyRobot - MagPuzzle
 - **Input:** mikrofon, webkamera och gaze tracker
 - **Output:** talande robothuvud
 - **Blickbeteende:** följer det som händer på bordet och användarens blickbeteende

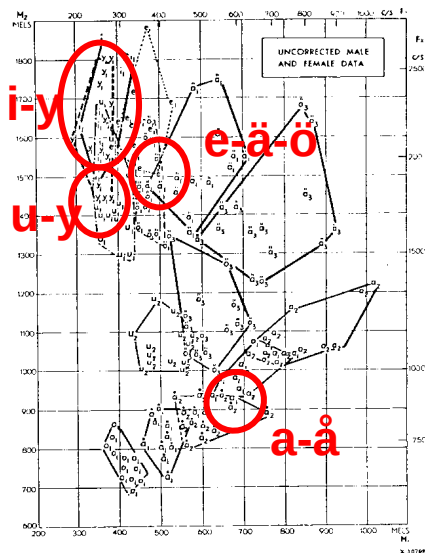
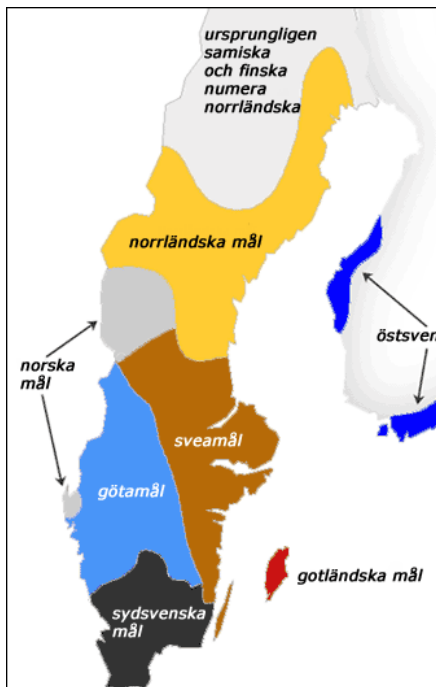




Varför är det svårt för maskiner att förstå mänskligt tal?

- Olika människor låter olika
- Samma person låter olika
- Bakgrundsljudet varierar

Skillnader mellan talare



- **Vi talar olika:** dialekt, sociolekt
- **Vi låter olika:** kön, storlek, ålder

Överlapp mellan talare

Människor kompenserar
automatiskt för detta
(men vi vet inte hur!)

Metoder krävs för
taligenkänning

Variationer för samma talare

Preposterous

BNP

YOLO!

RIP

NGO

ASAP

Levla

Helt
random



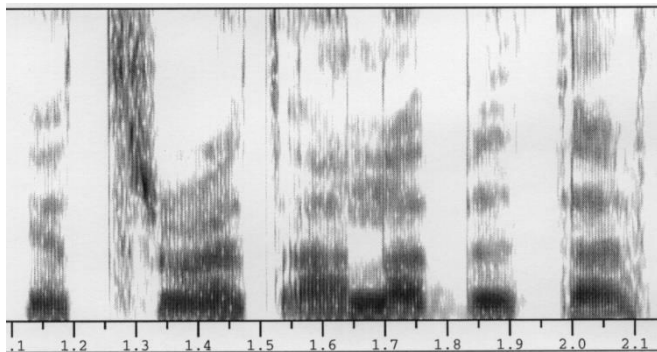
Beroende på sammanhang:
Formellt/informellt

Lyssnare (kompis, ålder,
ursprung, hörsel, ...)

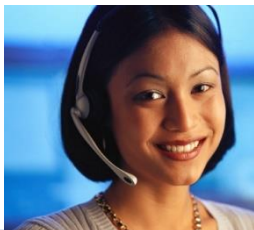
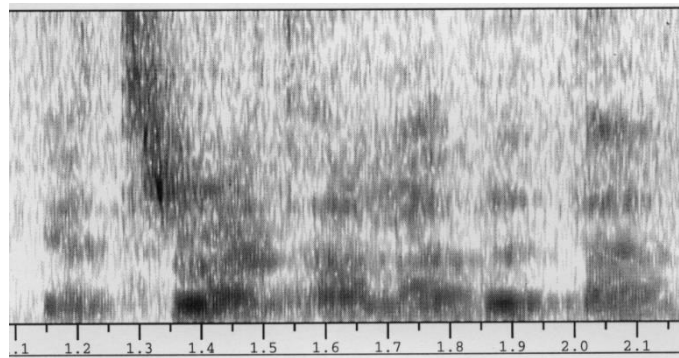
Känslor
Hälsa

Variationer i rumsakustik och bakgrundsljud

Headset



2 meter



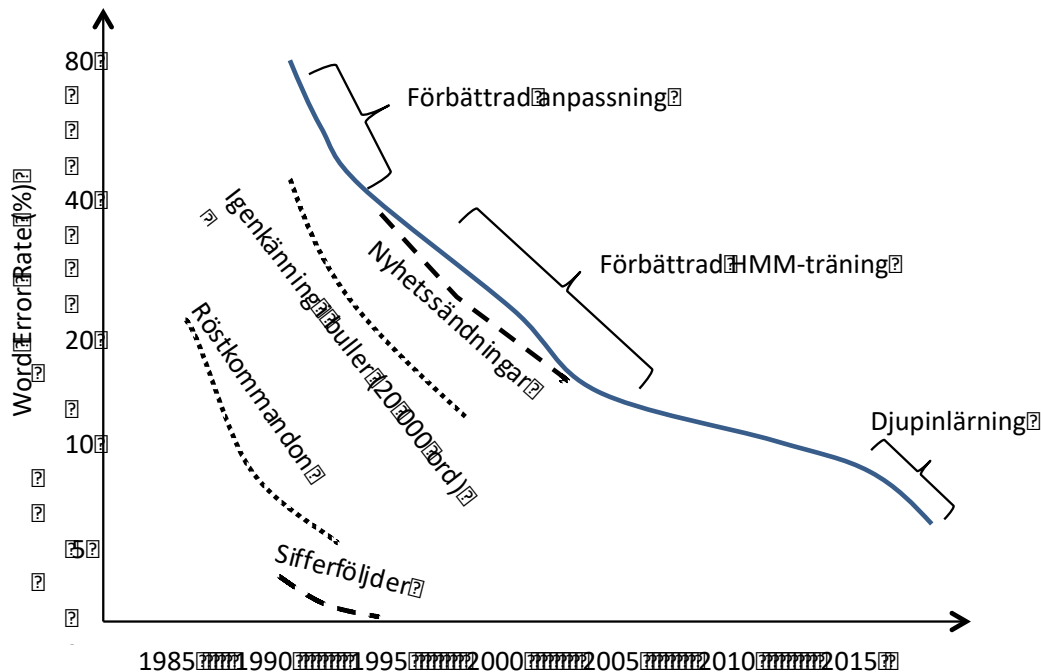


Vägen mot maskiner som förstår mänskligt tal

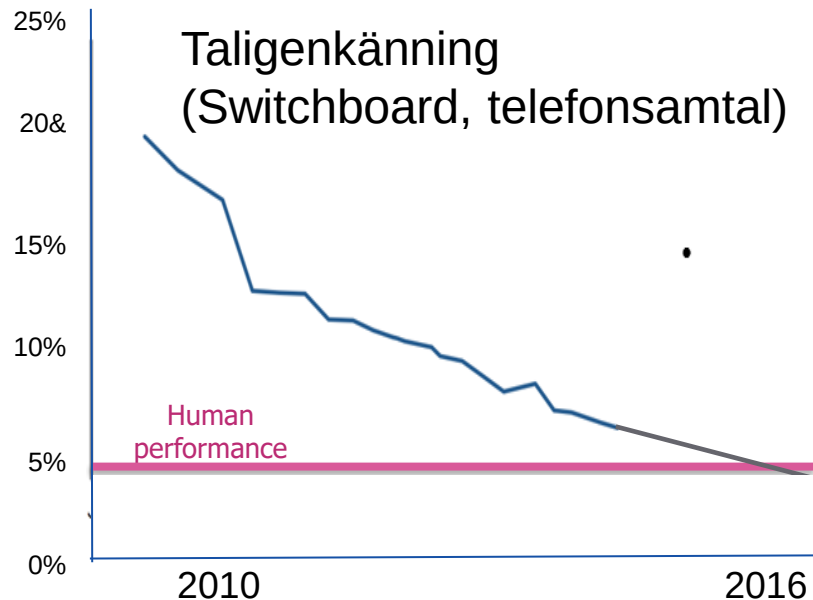
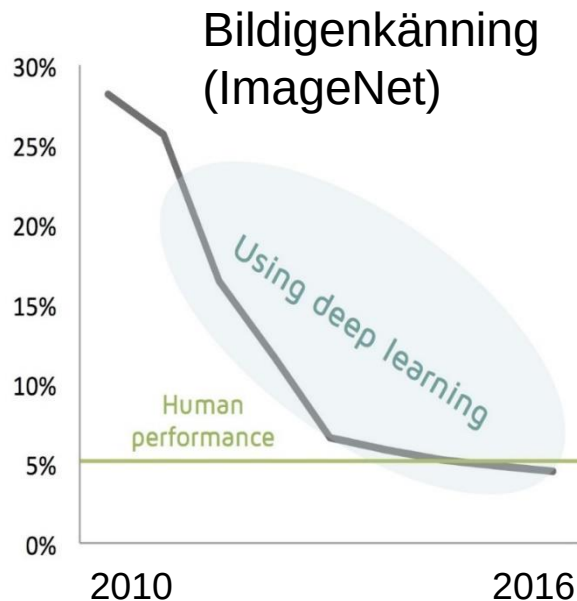
- 1952: Bell labs Audrey kunde känna siffrorna 0-9.
 - 1962: IBMs Shoebox maskin kunde känna igen 16 ord
 - 1971: CMUs Harpy kunde känna igen 1001 ord
 - 1980-talet fonem och HMMer -> tusentals ord eller mer
 - 1990-talet Dragon dictate känna igen tiotusentals ord i meningar
 - 2000-talet Google, SIRI och Nuance släpper taligenkänning i molnet för mobiltelefoner. De har slutat räkna ord, kan lägga till nya vid behov. Kvalitén förbättras dagligen
-

Utvärdering av taligenkänning

Räkna andelen feligenkända, missade, extra ord



Med hjälp av djupinärning har automatisk igenkänning av både bilder och tal lösts





Men maskiner har fortfarande svårt att förstå innebörden av vad folk säger?

- Människor säger inte alltid vad de menar rakt ut:
 - Vet du vad klockan är?
 - *Ja
 - Ibland säger människor motsatsen till vad de menar
 - Jag råkade sätta i batterierna fel i klockan
 - Det var ju smart gjort
 - Människor använder sig av kontexten när de talar:
 - Vad ska jag göra med klockan?
 - Ta den och ge till henne
-



Dessutom bestäms betydelsen av hur man uttalar orden

- Har Lisa klippt gräset? (Neutral fråga, ja eller nej.)
 - *Har* Lisa klippt gräset? (Förvånad, trodde inte hon hade gjort det)
 - Har Lisa *klippt* gräset? (Det var ju *vattna* det hon skulle göra.)
 - Har *Lisa* klippt gräset? (Det var ju *Pelle* som skulle göra det.)
 - Har Lisa klippt *gräset*? (Det var ju *häcken* hon skulle klippa.)
-



Fokus i detta föredrag är inte hur maskiner kan tänka som människor...

..utan snarare hur de kan tala som folk!

- Om man begränsar uppgiften är graden av förståelse som behövs för att lösa den via tal möjlig för en maskin
 - En telefonist i kundtjänsten kan till exempel koppla samtalet till rätt expert utan expertkunskap
 - Men om telefonisten talar otydligt eller väntar mycket länge innan hen svarar bryter kommunikationen ihop
 - Detsamma gäller talande maskiner
-

Talsyntes

Textuppläsning				Dialoger		
Regelbaserad syntes				Mönstermatchning		Statistikbaserat
Formantsyntes				Konkatenerings-syntes		Djupinlärning
Artikulatorisk syntes				Di- och trifoner	Unit selection	HMM
1950	1960	1970	1980	1990	2000	2010
						Neuronnät



Matematiska regler
beskriver hur olika ljud
påverkar varandra



Ljuddelar sparas i en
databas och sätts ihop
vid syntes



Matematisk
beskrivning,
statistiskt inlär

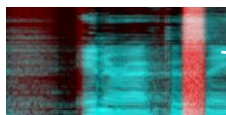
Talsyntes tränad på spontana samtal

Monoinspelning



27 h 2-
parts-
samtal

Andnings-
detektor



Taligenkänning

9 h tal från
ena talaren



Ord-
transkription

Disfluens-
detektor

um errr ahh eh
umm hah eh
h um errr uh
um uh uh

Disfluens-
annotering



Fluent
corpus



Disfluent
corpus



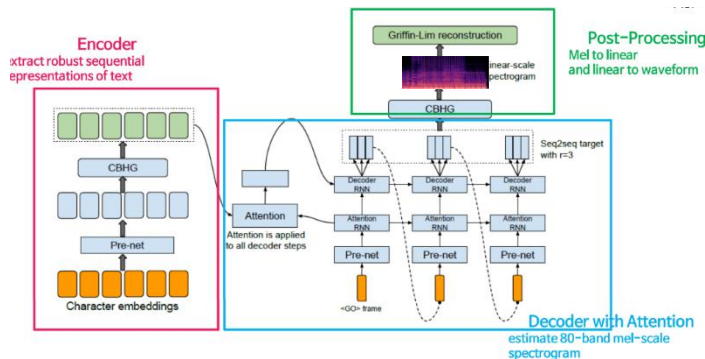
Läst tal-
korpus



Fluent pod-
korpus



Disfluent
podkorpus

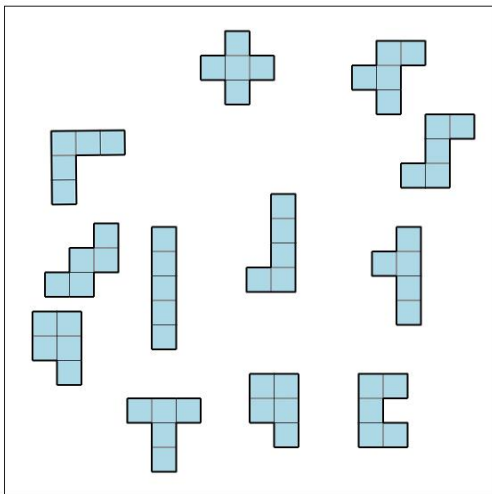


Székelly, E., Eje Henter, G., Beskow, J. and Gustafson, J. (2019) "Spontaneous Conversational Speech Synthesis from Found Data", In Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019), September 15-19, Graz, Austria.

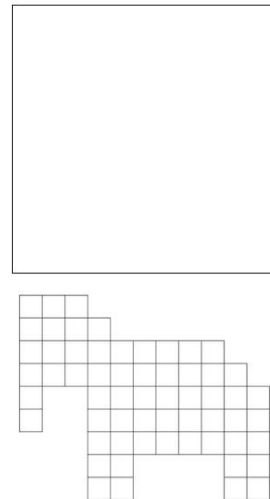
Resultat: en röst där vi kan styra hur den pratar

	Låg tonhöjd	Hög tonhöjd
Fluent, långsamt tal		
Initial eh, snabbt tal		
Medial eh , medelsnabbt		

Instruktioner genererade som läst tal jämfört med vår spontana röst



0%

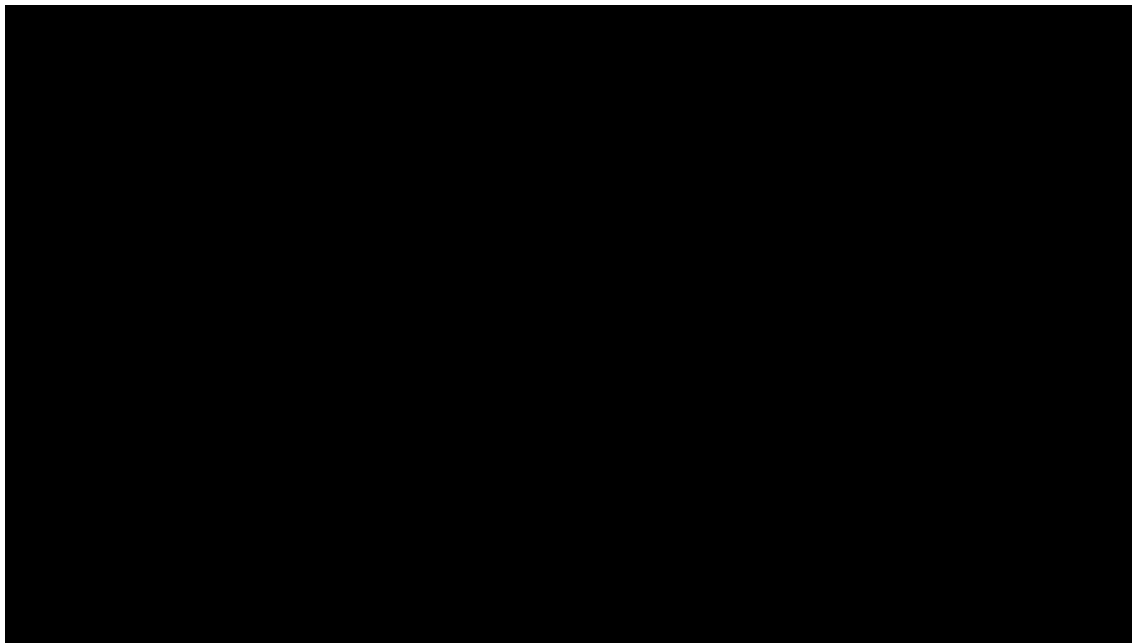


- 1 Nuance, Malcolm xpremium voice
- 2 KTH, Liam voice

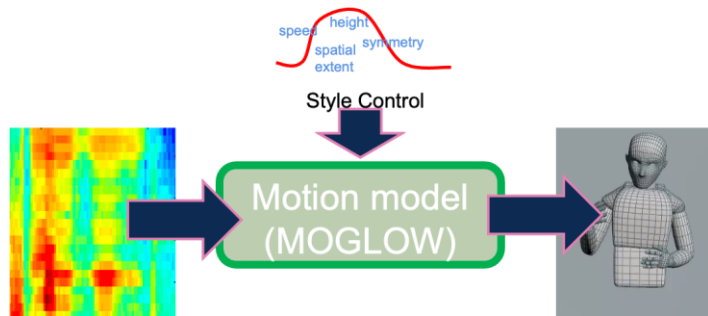




Uppläsning av undertexter översatta från franska



Generering av både tal och gester



The following clip is
generated from TEXT only



Att bygga en samtalssyntens

Vad lär vi oss från att använda riktiga data?

- Ordval, uttal och prosodi
 - Tvekljud
 - Förstärkningsord
 - Tilläggsfraser
 - Återkoppling
-

En svensk samtalsröst med personlighet

Tränad på 10 timmar från David Sundins podcast 80 väldigt goda mackor



Variationen i Davids spontanta tal

- Dialekt
- Engelska
- Crazy talk
- Läst tal
- Ljudeffekter
- Tal med mat i mun
- Engagerat tal
- Skrattande tal
- Mer skrattal
- Snabbtal
- Sång





Var han använder förstärkande ord

behöver ju

blir ju

ser ju

ska ju

den/det/man ju

har/hade ju

kan/kunde ju

kommer ju

är/var ju



Typiska ställen där David tvekar

..att eh..
..blev eh..
..den/det eh..
..då eh..
..därför eh..
..eller eh..
..en/ett eh..
..hur eh..
..i eh..
..och eh...
..på eh..
..ska eh..
..som eh..
..till eh..



Fraser David lägger till i slutet av meningar

..alltså
..då rå
..liksom
..eller nåt
..tror jag
..så att säga
..skulle jag säga
..såklart
..tycker jag



Återkoppling när David lyssnar

ja
mm
näe
jaha
ehm
nej
mhm
precis
aha
okej
va
oj
exakt

Vem är den riktiga David i denna dialog?



50 år av hjälpmedelsteknologi

60-talet

- Multimodala talskommunikationshjälpmedel (taktila & visuella)

70-talet

- Läsand maskiner för blinda
- Talprotes
- Talträning

80-talet

- Symbolbaserat kommunikationshjälpmedel
- Talstyrning i hemmet

90-talet

- Personlig talprotes
- Läppsynkad avatar som hörstöd

00-talet

- Ordprediktion för dyslektiker
- Avatar som påminnare



På senare har vi börjat forska på sociala robotar för hälsa och välmående



- **Kognitiv nedsättning**
 - Detektion (dementia)
 - Rehabilitering (stroke)
- **Självständigt boende för äldre**
 - Påminnare (medicin, träning)
 - Fysisk hjälps (tex till toaletten)
 - Sällskap (spel, telepresence)
 - Motivationscoach (äta träna)

EACare: detektion av demens

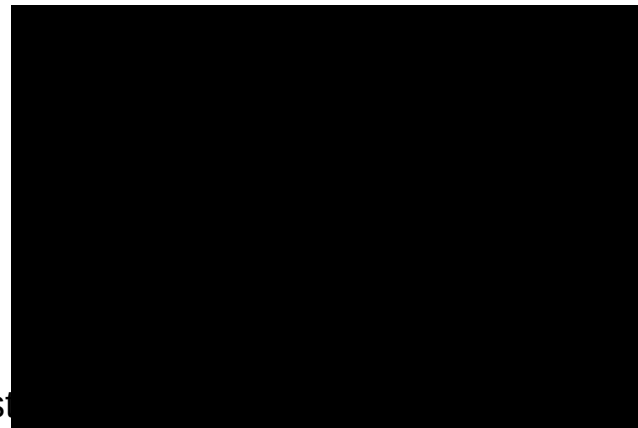


Steg 1 – datainsamling

- 100 interaktioner läkare-patient
- minnestest
- Ca 1 timme per session
- >10 synkroniserade dataströmmar

Step2 – utveckling

- detektion of biomarkörer för demens
- social robot som kan utföra minnestest





Nuvarande pilotprojekt: bygga en röst för en ALS-sjuk

- **Nuvarande lösning:** blickstyrning, genererell kommersiell röst
- **Problem:** “Jag vill inte låta som försäkringskassan, och jag vill låta som om jag menar vad jag säger”
- **Vår lösning:** Vi spelar in två systrar och bygger en spontan talsyntesröst

Krav på en röst för kommunikationshjälpmedel

- *Jag behöver teknologi som fångar hur jag verkligen pratar i en konversation.*
- *Om jag får säga det själv är jag en nyanserad och humoristisk person och jag vill att min syntetiska röst ska avspegla detta.*





Vad finns kvar att forska på?

- **Taligenkänning:**

- Olika uttal (inte modersmål, barn)
- Olika talstilar (barn, äldre, ovana användare, emotioner)
- Naturlig konversation

- **Talsyntes:**

- Varierad talstil (kontrollera betoning, känslor)
- Andra språkljud (återkoppling, skratt)
- Namn, utländska ord

- **Dialogsystem:**

- Social konversation
 - Bredare ämnen
 - Sunt förnuft
 - Längre minne
-

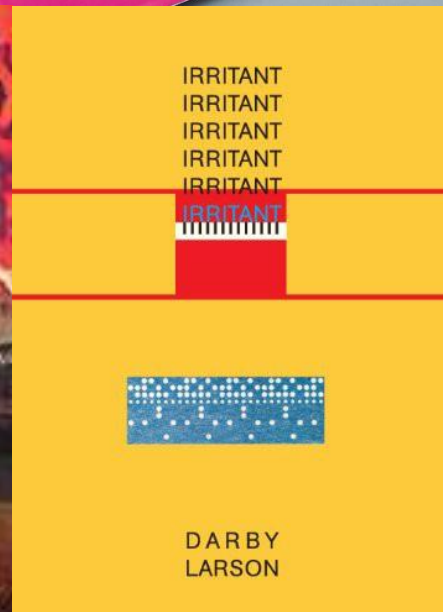
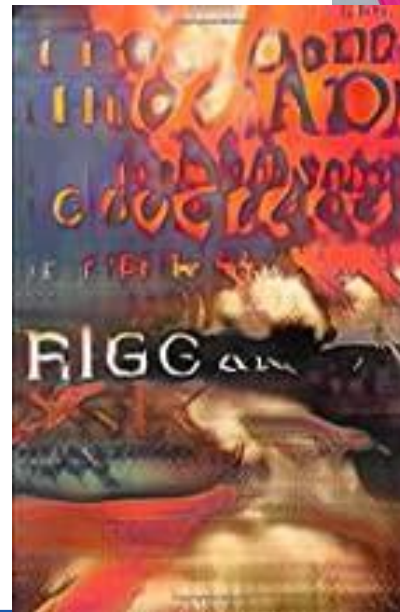


Kan maskiner vara kreativa?

- Litteratur
- Konst
- Musik
- Dans

Böcker skriva av AI

- 1 the Road — Ross Godwin and Kenric McDowell
- The Day a Computer wrote a Novel
- Dinner Depression — Julia Joy Raffel
- World Clock — Nick Montfort
- Irritant by Darby Larson





AI kan generera nya citat

If I'd known I was going to have to rely on the kindness of strangers, I would have bought a gun.

Groucho Marx 2.0

Too much sanity is as bad as too little sanity.

Oscar Wilde 2.0

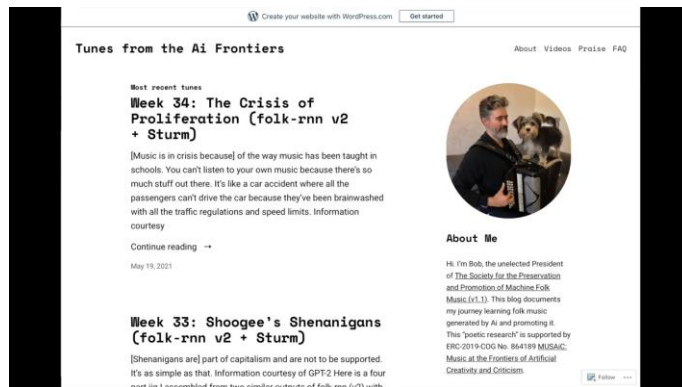
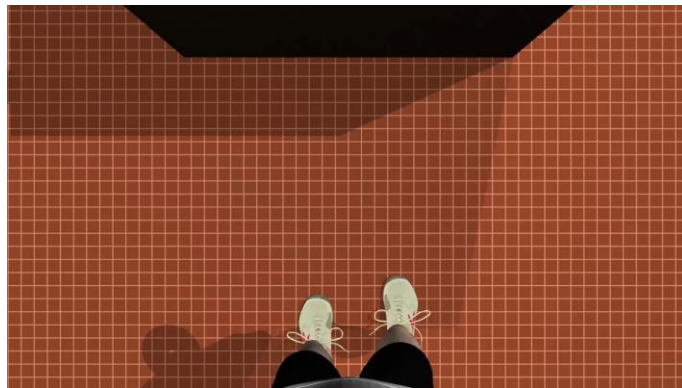


AI kan generera en beskrivning anpassad till en viss målgrupp

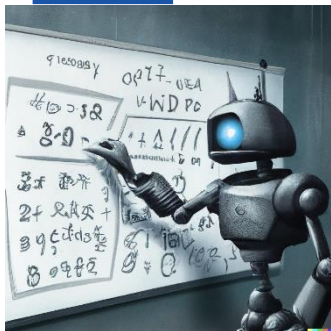
- **Explain the moon landing to a professor old in a few sentences.**
 - America landed 12 astronauts on the surface of the Moon from 1969-1972. The first landing was on July 21, 1969, and the last on December 14, 1972. The astronauts spent a total of about 28 hours on the Moon.
 - **to a social worker**
 - The moon landing was a major landmark in human history that helped push the boundaries of human achievement.
 - **to a teenage girl**
 - The moon landing is when astronauts walked on the moon and took photos of the moon.
 - **to a 6 year old**
 - A group of people went up in a spaceship and landed on the moon. They collected some rocks and put them in a bag.
-

Musik skriven av datorer

- Songularity
 - AI tränad på Morrissey's låttexter och recensioner av workout-DVDn P90X.
- FolkRNN
 - Bob Sturm på KTH genererar nya irländska folkmusikmelodier



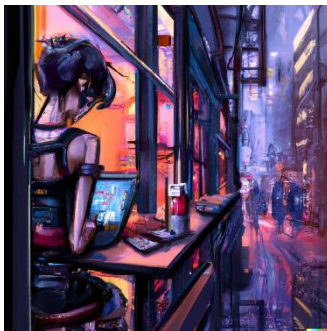
Bilder skapade från text



A robot attempting to learn a new language



A fish fishing



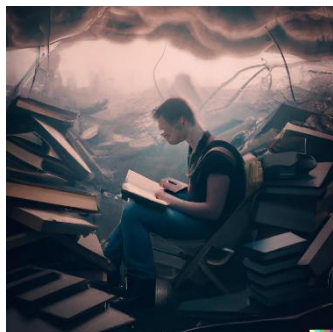
Cyberpunk illustration of girl working in cafe in NYC



Eye of the sunflower



A raccoon astronaut with the cosmos reflecting on the glass of his helmet dreaming of the stars, digital art



Reading books during Apocalypse



Steampunk honeybees with headsets harvesting pollen



Two little girls traveling through space on bikes



A hand-drawn sailboat circled by birds on the sea at sunrise



A medieval painting of the wifi not working

Musikvideo automatiskt genererad från låttext



VQGAN/CLIP med hjälp av Jonas Beskow

Dans AI-genererad från musiken tränad på människor som dansar

