Speech Technology – An Introduction

Robert Eklund

robert@roberteklund.info
http://roberteklund.info

University Lecturer (Phonetics) at
Department of Culture and Communication, Linköping University
Linköping, Sweden
&
Affiliated Researcher (Neurocognition) at
Department of Neuroscience / Karolinska Institute
Stockholm, Sweden
&
Associate Professor/Docent (Computational Linguistics) at
Department of Computer Science, Linköping University
Linköping, Sweden

30 September 2012

This Talk

- What is language (short intro)
- What is speech technology?
- Different areas
- Why speech technology?
- Problems?
- History
- Future…

< PICTURE HERE >

Lt Commander Data in Star Trek

Different Areas

- Automatic Speech Recognition (ASR)
- Speech Synthesis
- Text-to-Speech Systems
- Speaker Verification / Identification
- Machine Translation
- Dialogue Systems
- Facial Animation
- Multimodal Systems
- Androids
- “Neurotechnology”

1. Introduction

Introduction (1)

- Origin of speech unknown
- Writing systems 5000 years old (Mesopotamia, Egypt, China)
- ... fully developed languages, speech much older
- Human brain/speech apparatus adapted to speech production
- Chimps *can’t* produce speech sounds
- Hoover the Talking Seal
  * 1971, Maine
  † 1985, Boston

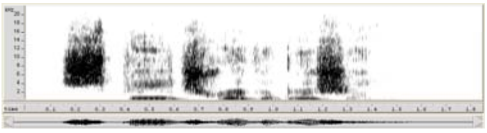
Introduction (2)

- There are around 5000–8000 languages
- Very uneven distribution
- English (often) listed as 2nd (Swedish 85th)
- Two billion people speak English each day
- Technological development → Speak with computers

2. Speech Recognition

Speech Recognition (1)

- Computers that “understand” (at least recognize) speech
- Spectrograph invented in 1946 → “read” speech



- First system described in Davis, Biddulph & Balashek (1952)

Speech Recognition (2)

- Variability much greater than previously assumed
- Variability (still) the major problem
- Speech vary as a function of (among other things)...
  - Gender
  - Age
  - Dialect
  - Sociolect
  - Individual (inter-/intra-)
  - Speech rate / reductions (“it is green” → “screen”)
  - Disfluencies (“uh”, “uhm” etc)

Speech Recognition (3)

- Speech recognizers are “trained”
- Create sets of sentences that cover language in question (phonetically and linguistically)
- Speakers record these sentences
- Speakers should represent variability as to gender, age, dialects, sociolects and so on
- Computer compares recordings with transcriptions
- Computer builds a model of variability
- But *what*, exactly, should be the training material?

Speech Recognition (4)

- Linguistic term: phoneme
- Definition: “Smallest meaning-carrying unit”

1. b/p changes the meaning in English, but not in Finnish

   Ex: *beer* vs *peer*

2. v/w changes the meaning in English, but not in Swedish

   Ex: *vie* vs *why*

- Number of phonemes varies between languages, from a low 11 (Rotokas) to 141 (!XŨ)
- English and Swedish both around 45

Speech Recognition (5)

- Two different terms (often confused):

  Speech recognition

  … recognizes *what* is said, irrespective of *who* says it

  Speaker recognition (identification / verification)

  … recognizes *who* speaks, irrespective of *what* is being said

- Children harder to recognize than women, who in turn are harder to recognize than men (diminishing problem, though)
- Acoustic reason for this:
  The higher the pitch, the greater the search space

3. Speech Synthesis

Speech Synthesis

- Computers that talk
- Three genuine forms:
  1. Articulatory synthesis
  2. Formant synthesis
  3. Concatenative synthesis
- Pseudo synthesis:

  "Canned Speech"

Canned Speech

- Record a carrier phrase:

  The train departs at _______ o'clock
- Fill empty slot with prerecorded material
- Each new utterance requires new recording, preferably with the same speaker (might have left the company)
- Partly good sound quality (recorded bits)
- Low flexibility!

Articulatory Synthesis (1)

- Attempts to mimic human speech behaviour
- Uses virtual…

  - lungs
  - vocal folds
  - cheeks
  - tongue
  - palate
  - air streams
  - lips
  - teeth

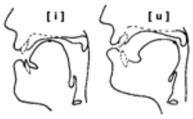
  … and so on and so forth

Articulatory Synthesis (2)

- Urban Hjärne (1641–1724)
- Used decapitated heads
- Bellows pumped air through throat
- Strings attached to tongue, lips etc
- Poor results
- Ethical problems

Articulatory Synthesis (3)

- Alexander Graham Bell (1847–1922)
- Used his (living) dog's throat
- Dog howled, Bell squeezed

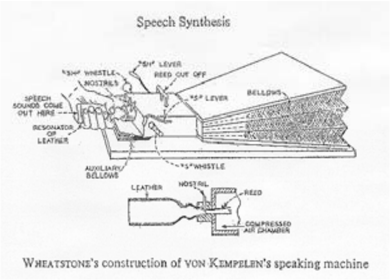


- Managed to produce vowels
- Poor results
- Ethical problems

Articulatory Synthesis (4)

- **Wolfgang von Kempelen (1734–1804)**
- **Speech machine (1791)**
- **Mechanical automaton**
- **Worked!!**

Articulatory Synthesis (5)



WHEATSTONE's construction of VON KEMPELEN's speaking machine

Thanks to Mária Gósy, Kempelen Farkas Speech Research Laboratory, Budapest.

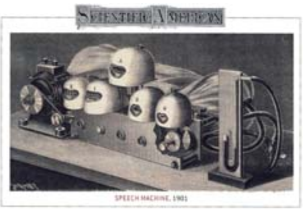
Articulatory Synthesis (6)

- **Kempelen already famous for chess-playing machine (1770)**
- **"The Turk".**
- **Poe fascinated**
- **Human hidden inside**
- **Bad "cred"**
- **Early example of artificial intelligence**

Articulatory Synthesis (7)

- **From Scientific American (1901).**

"Dr. Marage has constructed an apparatus, using the plastic substance employed by dentists, to reproduce the interior of a person's mouth while pronouncing the different vowels."



SPEECH MACHINE, 1901

Articulatory Synthesis (8)

- **Hideyuki Sawada, Kagawa University.**

< FILM CLIP HERE >

http://www.eng.kagawa-u.ac.jp/~sawada/index_e.html

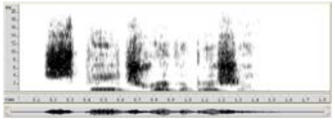
Articulatory Synthesis (9)

- **Difficult!**
- **From a research perspective most interesting**
- **From a commercial perspective not so interesting (so far)**
- **If successful it can do everything humans can do**

Formant Synthesis (1)

- Imitates acoustics of speech
- Formants
  = frequency bands
  with high amplitude
  in speech signal

  = “tube resonances”



- Pitch = melodic contour = Formant 0 = F0
- Generate tone pulse in computer
- Filter tone electronically until it sounds like speech

Formant Synthesis (2)

- Early synthesizer:

  The Voder (1939)

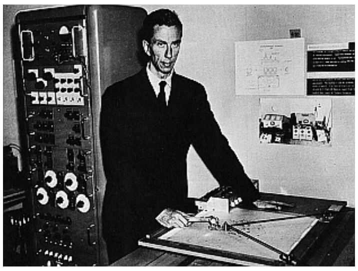
- Homer Dudley

Formant Synthesis (3)

- Early Swedish synthesizer (that speaks English):

  Ove (1953)

- Gunnar Fant



Concatenative Synthesis (1)

- “Cut-and-paste” synthesis
- Record real/authentic human speech
- Use “sound pieces” (cut)

1. Phonemes:	Impossible! Not same [k] in ”*cat*” and ”*kit*”
2. Diphones:	/ki/ka/ku/ko etc
3. Polyphones:	spri-/-orv/-olmskt etc
4. Arbitrary units:	*Unit selection*

- Assemble in new orders (paste)

Concatenative Synthesis (2)

- Recording:

	∅	_ta	IGEN
	Säga	tat	IGEN
	Säga	tal	IGEN
	Säg	alsa	IGEN
	Säga	språt	IGEN
	Säga	tåk	IGEN
	SÄGA	tak__	∅

- Synthesis: t + ta + al + ls + språ + åk + k
  (“spoken language”)

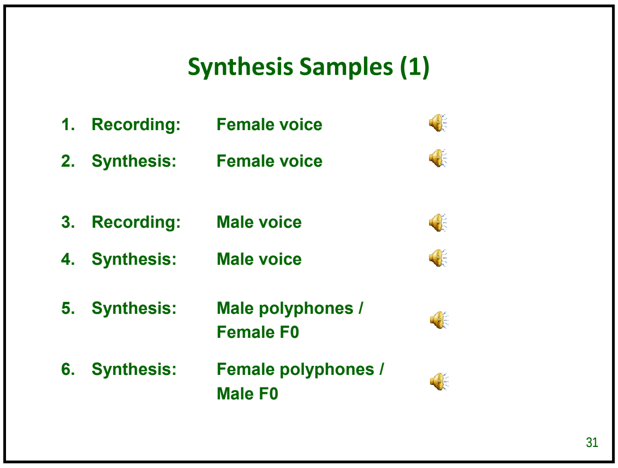
Concanenative Synthesis (3)

- Number of polyphones varies between languages
- *Phonotactics*: How phonemes can be combined
- Swedish allows many different combinations
- Typical syllable structure: CV or V syllables
- Swedish maximum: CCCVCCCCCCCCC

  Ex: CCV … VCCCCCC

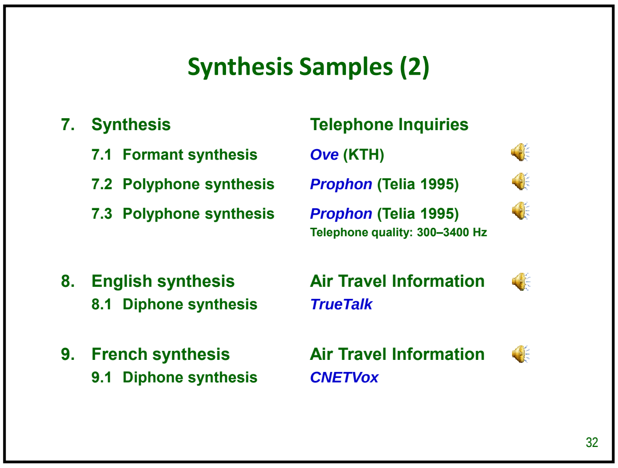
  ”*… ordet stockholmskts … uttal*”

- Good quality!
- Commercially interesting, especially *unit selection*



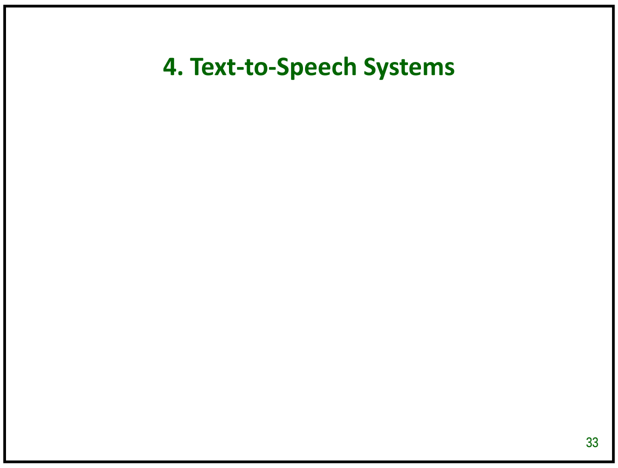
Synthesis Samples (1)

1.	Recording:	Female voice	🔈
2.	Synthesis:	Female voice	🔈
3.	Recording:	Male voice	🔈
4.	Synthesis:	Male voice	🔈
5.	Synthesis:	Male polyphones / Female F0	🔈
6.	Synthesis:	Female polyphones / Male F0	🔈

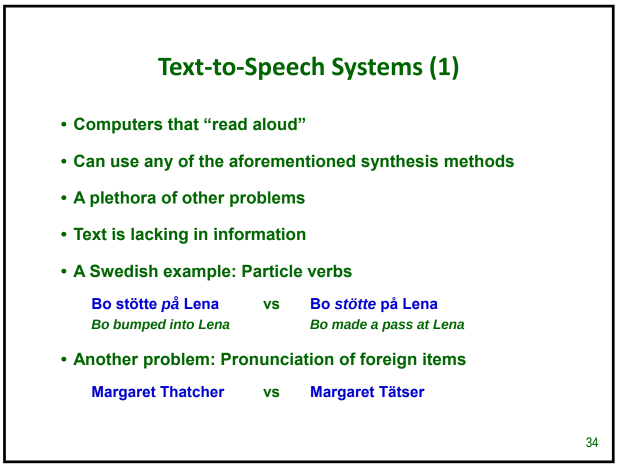


Synthesis Samples (2)

7.	Synthesis	Telephone Inquiries	
7.1	Formant synthesis	*Ove* (KTH)	🔈
7.2	Polyphone synthesis	*Prophon* (Telia 1995)	🔈
7.3	Polyphone synthesis	*Prophon* (Telia 1995) Telephone quality: 300–3400 Hz	🔈
8.	English synthesis	Air Travel Information	🔈
8.1	Diphone synthesis	*TrueTalk*	
9.	French synthesis	Air Travel Information	🔈
9.1	Diphone synthesis	*CNETVox*	



4. Text-to-Speech Systems



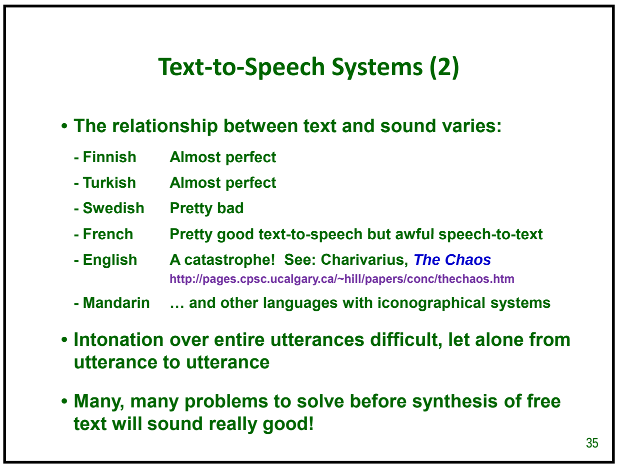
Text-to-Speech Systems (1)

- Computers that “read aloud”
- Can use any of the aforementioned synthesis methods
- A plethora of other problems
- Text is lacking in information
- A Swedish example: Particle verbs

Bo stötte *på* Lena *Bo bumped into Lena*	vs	Bo *stötte* på Lena *Bo made a pass at Lena*

- Another problem: Pronunciation of foreign items

Margaret Thatcher	vs	Margaret Tätser

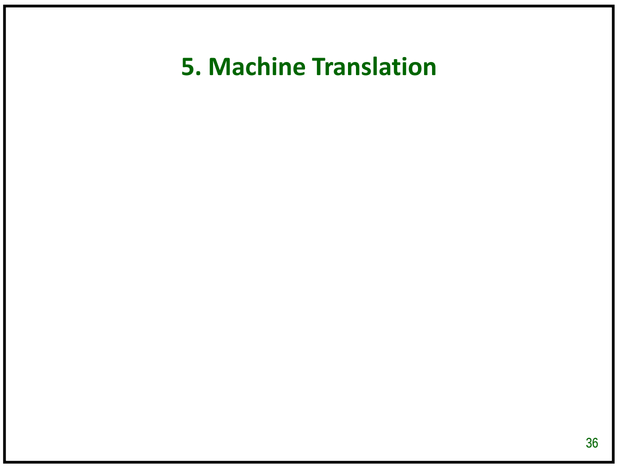


Text-to-Speech Systems (2)

- The relationship between text and sound varies:

- Finnish	Almost perfect
- Turkish	Almost perfect
- Swedish	Pretty bad
- French	Pretty good text-to-speech but awful speech-to-text
- English	A catastrophe! See: Charivarius, *The Chaos* http://pages.cpsc.ucalgary.ca/~hill/papers/conc/thechaos.htm
- Mandarin	… and other languages with iconographical systems

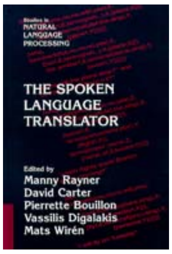
- Intonation over entire utterances difficult, let alone from utterance to utterance
- Many, many problems to solve before synthesis of free text will sound really good!



5. Machine Translation

Machine Translation (1)

- Automatic translation between languages
- Text-to-Text or Speech-to-Speech
- Project: Spoken Language Translator
  - Telia Research
  - SRI International, Menlo Park, California
  - SRI International, Cambridge, UK
- 1993–1999
- Air Travel Information Service (ATIS)
- English/Swedish/French
- Cambridge University Press



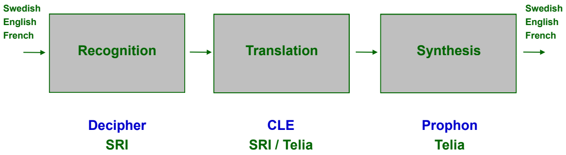
Studies in NATURAL LANGUAGE PROCESSING

THE SPOKEN LANGUAGE TRANSLATOR

Edited by
Manny Rayner
David Carter
Pierrette Bouillon
Vassilis Digalakis
Mats Wirén

Machine Translation (2)

Spoken Language Translator



Swedish
English
French

Recognition → Translation → Synthesis

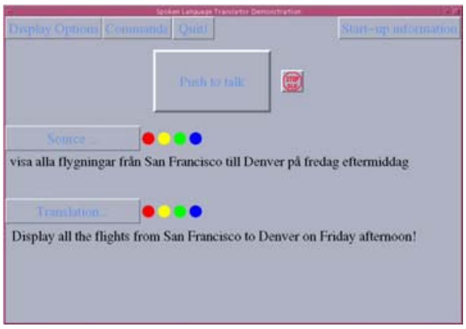
Swedish
English
French

Decipher
SRI

CLE
SRI / Telia

Prophon
Telia

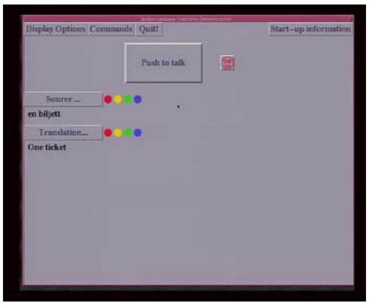
Machine Translation (3)



visa alla flygningar från San Francisco till Denver på fredag eftermiddag

Display all the flights from San Francisco to Denver on Friday afternoon!

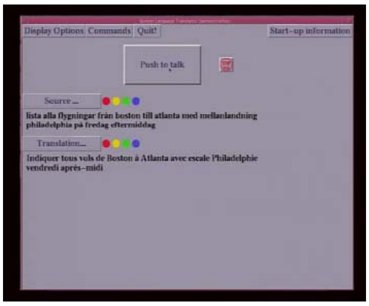
Machine Translation (4)



en biljett

One ticket

Machine Translation (5)



lista alla flygningar från boston till atlanta med mellanlandning philadelphia på fredag eftermiddag

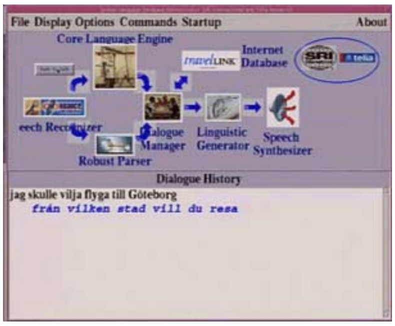
Indiquer tous vols de Boston à Atlanta avec escale Philadelphie vendredi après–midi

6. Dialog Systems

Dialog Systems (1)

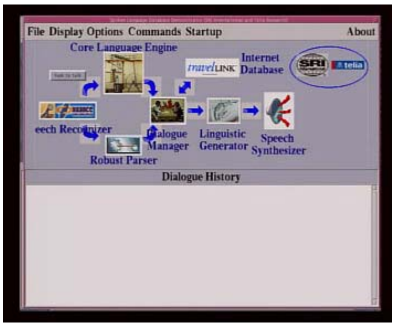
- Systems that both listen and talk
- Communication with e.g. databases
- Requires other kinds of linguistic knowledge
- Conversation grammars
- Dialog management (grammars)
- Commercially interesting

Dialog Systems (2)



jag skulle vilja flyga till Göteborg
från vilken stad vill du resa

Dialog Systems (3)



7. Facial Animation

Facial Animation (1)

- Speech synthesis with a face
- Improves understanding of speech
- McGurk effect
- Everyone “listens” with their eyes
- Original reference:

Harry McGurk & John MacDonald. 1976.
Hearing lips and seeing voices
Nature, vol. 264, pp. 746–748.

Facial Animation (1)

< FILM CLIP HERE >

http://www.youtube.com/watch?v=aFPtc8BVdJk

Facial Animation (3)

- Different methods
- Telia Research: own/unique method
- Same principle as for speech concatenation
- Don't remember? OK...

49

Facial Animation (4)

- Recording:

∅	_ta	IGEN
Säga	tat	IGEN
Säga	tal	IGEN
Säg	alsa	IGEN
Säga	språt	IGEN
Säga	tåk	IGEN
SÄGA	tak_	∅

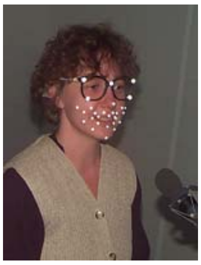
- Synthesis: t + ta + al + ls + språ + åk + k
  ("spoken language")

… plus animated face

50

Facial Animation (5)

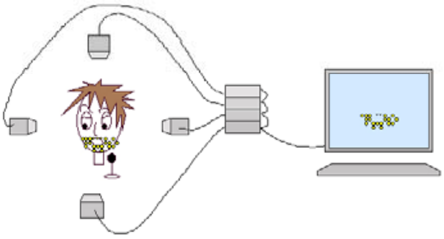
- Record facial movements at the same time as sound
- Reflectors in face of speaker
- 24 reflectors around the mouth (and nose)
- Glasses to normalize for head movements



51

Facial Animation (6)

- Recording:



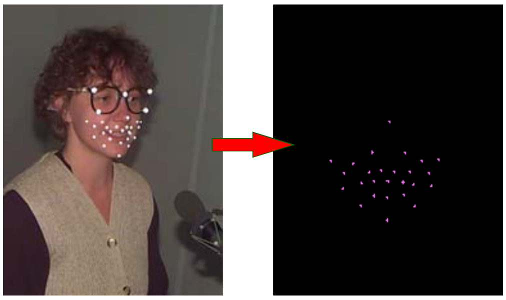
52

Facial Animation (7)

- Laboratory setting

53

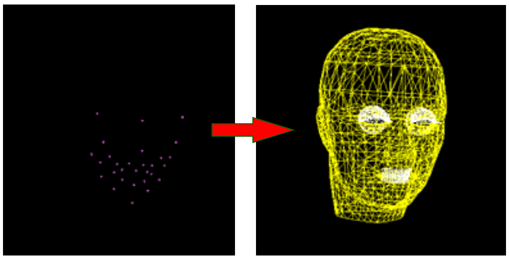
Facial Animation (8)



Reflector movements stored as… … 3D-movements in computer

54

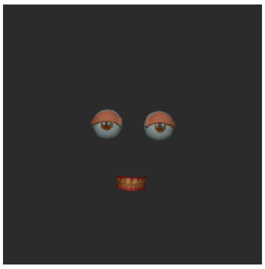
Facial Animation (9)



Reflectors attached to… … wire model of head

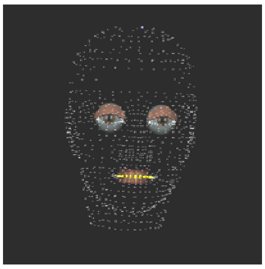
Facial Animation (10)

- Add eyes, teeth...
- Computer graphics



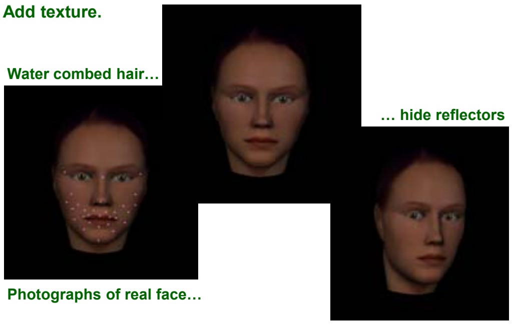
Facial Animation (11)

- Head shape:



Facial Animation (12)

- Add texture.



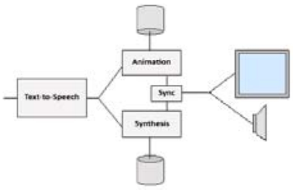
Water combed hair…

… hide reflectors

Photographs of real face…

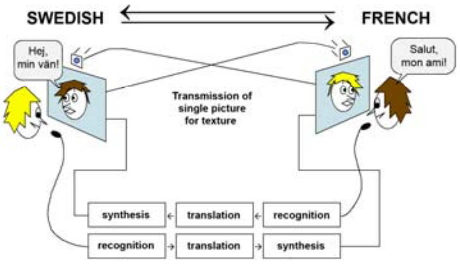
Facial Animation (13)

- System:



Facial Animation (14)

- Vision:



Facial Animation (15)

Sign up with [illegible] get a discount.

8. Multimodal interfaces

Multimodal interfaces (1)

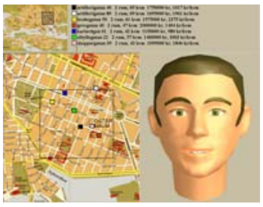
- **Combination of modalities**
  - **- Speech**
  - **- Computer mouse**
  - **- Gestures**
  - **- Facial expressions**
  - **- Touch**
  - **- Smells**
  - **… etc**
- **Both for input/user and output/computer**

Multimodal interfaces (2)

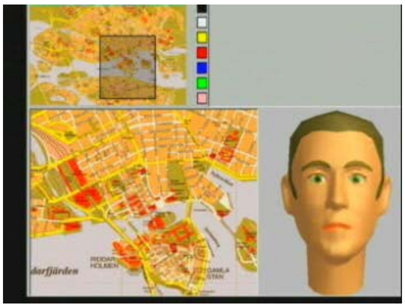
- **How do humans interact with machines/computers?**
- **Reeves & Nass (1996):**
  The Media Equation
- **"It's Only A Movie" phenomenon.**
- **Nass, Kim & Lee (1998):**
  When Your Face is the Interface: An experimental comparison of interacting with one's own face or someone else's face
- **Jönsson & Dahlbäck (1988):**
  Talking to your computer is not like talking to your best friend

Multimodal interfaces (3)

- **AdApt.**
- **Telia and KTH**
  (Royal Institute of Technology, Stockholm)
- **Apartment info**



Multimodal interfaces (4)



[illegible]
[illegible] olm.

Multimodal interfaces (5)

- Waseda University (Japan): Humanoid
- Conversational robot
- Integrates:
  - Speech recognition
  - Speech synthesis
  - Dialog grammars
  - Computer Vision
  - ”Gaze Tracking”
  - Functioning robotics (autonomous)

Multimodal interfaces (6): Humanoid

< FILM CLIP HERE >

http://www.humanoid.waseda.ac.jp/

Multimodal interfaces (7): Aiko

< FILM CLIP HERE >

http://www.aikoproject.jp

http://www.youtube.com/watch?v=lg7qmddOq4s&feature=related

9. “NeuroTechnology”

“NeuroTechnology”

- Not speech technology in the strict sense
- However, language/speech often involved
- Brief (!) summary of some areas

Brain–Reading

- Via ”neuroimaging” directly read brain activity
- EEG: Net of electrodes to measure electrical activity
- fMRI: Measure oxygen consumption (assumed to reflect activity)
- Via signals one has succeed to detect what:
  - … kind of image subjects look at (face, cat, chair etc) (Haxby et al., 2001)
  - … image out of 1000 images a person is looking at (Kay et al., 2008)
  - … scene in Sergio Leone’s movie *The Good, The Bad and The Ugly* a person is watching (Hasson et al., 2004)

< PICTURE HERE >

Brain–Computer Interfaces (BCI)

- Interface directly between a brain and a computer
- Net of electrodes registers electrical activity in the brain which is used to control a cursor
- Application area: “locked-in”-patients (quadriplegics)

< PICTURE HERE >

- “Neural prosthetics” (Andersen et al., 2004)

http://computer.howstuffworks.com

Lie Detection

- Normal lie detectors not reliable or allowed in courts
- Hypothesis: different brain activity associated with truths and lies
- Method: via EEG or fMRI directly investigate whether a specific brain produces a truth or a lie
- Hot area (duh!)

< PICTURE HERE >

http://singularityhub.com/2010/05/06/another-attempt-to-use-fmri-lie-detector-in-us-court-fails-in-brooklyn-more-on-the-way/

- ”Not there yet”

10. Summing Up

Summing Up

- Lots of specific problems to solve
- General technology development creates new challenges
- Several different and ambitious targets
- Several application areas possible
  - People with disabilities
  - Company
  - News anchors (?)
- Language plays a crucial part when creating androids
- … and don’t forget ”The Turk”!

Fiction vs Reality

- So, how far have we come?
- … well, just compare:

< PICTURE HERE >

Lt Commander Data

< PICTURE HERE >

Aiko

Summing it up

… and finally a comment, “as seen on TV”…

TV3 Direkt, 3 November 1998

[illegible] Telia believes that we will have such faces in our cell-phones already within a few years, to read out e-mail messages. [illegible] seconds later reads out a translation. [illegible] you prefer to have a little "person" who lives in it, like "Anna" in this case, with whom you can communicate.

Thanks for listening!

Contact me:

robert@roberteklund.info

Homepage:

http://roberteklund.info