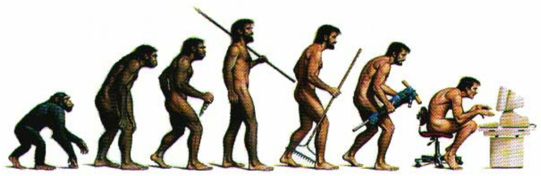
Human-Computer Interaction

Termin 9:

Spoken Language Interaction

The evolution of user interfaces (and the rest of this lecture)



Year	Paradigm	Implementation
1950s	None	Switches, punched cards
1970s	Typewriter	Command-line interface
1980s	Desktop	Graphical UI (GUI), direct manipulation
1980s+	Spoken Natural Language	Speech recognition/synthesis, Natural language processing, dialogue systems
1990s+	Natural interaction	Perceptual, multimodal, interactive, conversational, tangible, adaptive
2000s+	Social interaction	Agent-based, anthropomorphic,social, emotional, affective, collaborative

Using *speech* to interact with systems

- Intuitive form of communication, no need for training
- Relates to (one) way of thinking; *but* images, maps, …
- Paradigm: Computer adapts to human way of interaction

Speech interaction

Used today

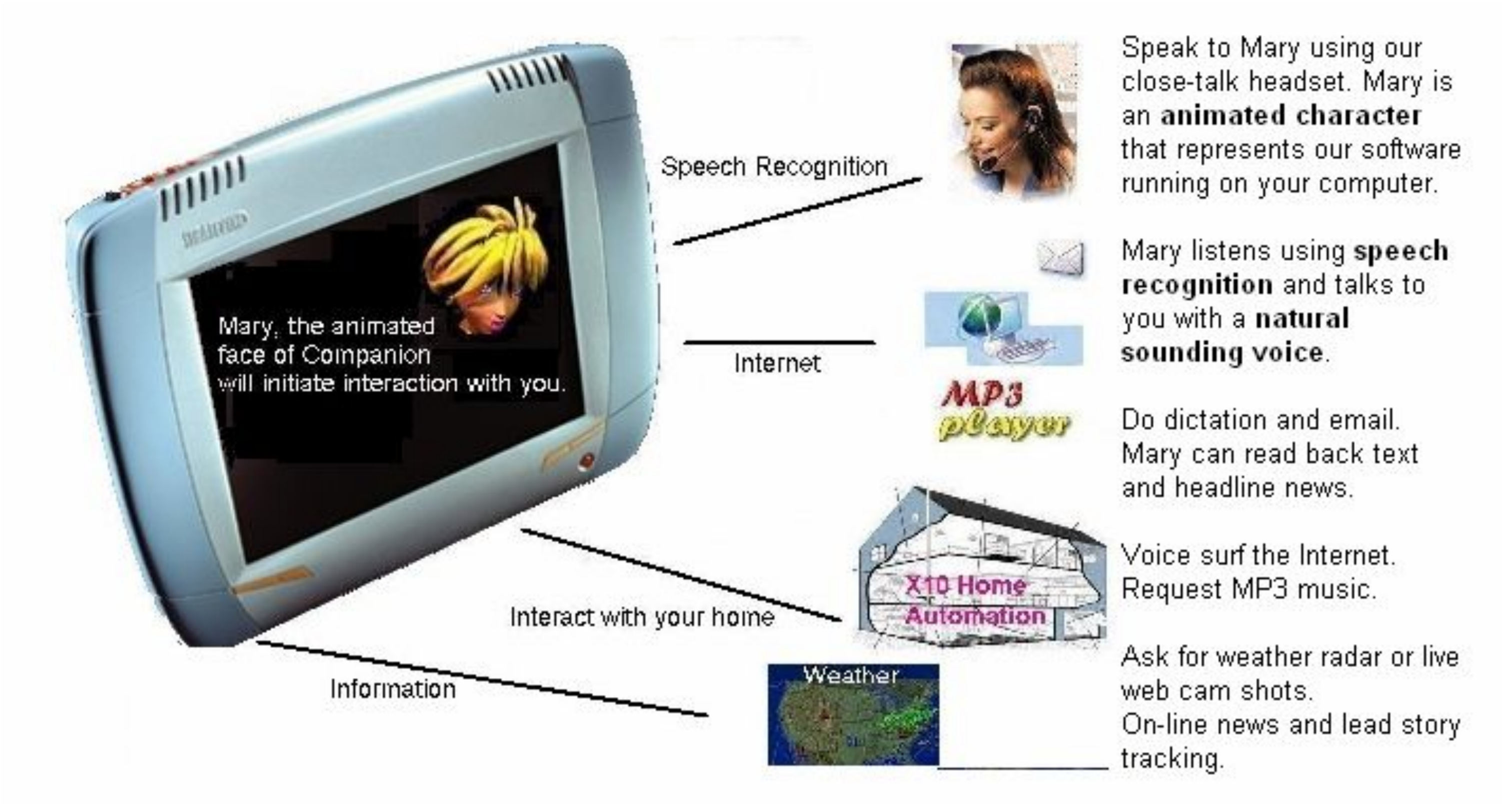
- on the desktop, e.g. dictate
- on the phone, e.g. ticket booking, pizza ordering

Research for

- mobile devices
- automotive interaction
- Virtual Reality
- conversational agents
- mobile robot companions

SmartKom

Cutting edge technology



http://www.talkingdesktop.com/concept.htm

Spoken Language Dialogue Systems (SLDS)

- A system that allows a user to *speak* his queries in natural language and receive useful spoken *responses* from it

- Provides an interface between the user and a computer-based application that permits *spoken interaction* with the application in a “relatively natural manner”

Levels of sophistication

- Touch-tone replacement:

  System Prompt: "For checking information, press or say one."
  Caller Response: "One."

- Directed dialogue:

  System Prompt: "Would you like checking account information or rate information?"
  Caller Response: "Checking", or "checking account," or "rates."

- Natural language:

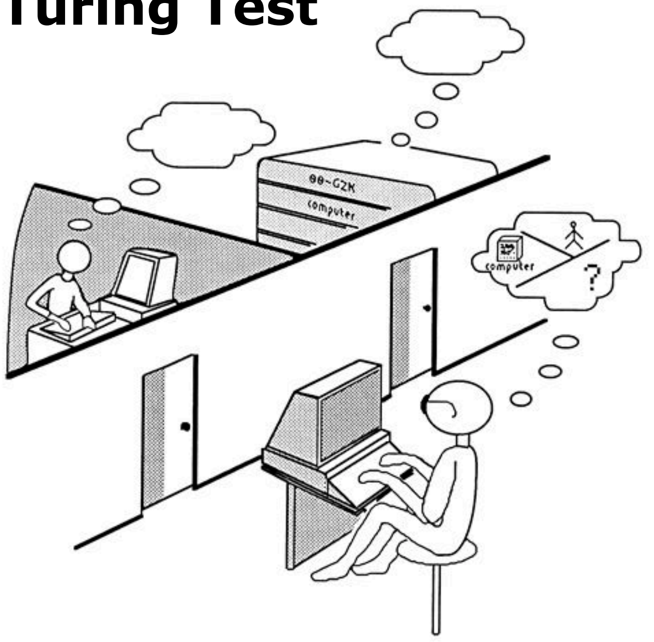
  System Prompt: "What transaction would you like to perform?"
  Caller Response: "Transfer 500 dollars from checking to savings."

Levels of sophistication

Controlled language	limited vocabulary, simple grammar (e.g. command language)
Natural language	huge vocabulary, complex grammar, grammatical variation, ambiguities, unclear sentence boundaries, omissions, word fragments
Natural dialogue	turn-taking, initiative switch, discourse grounding, restarts, interruptions, interjections, speech repairs

Perfect natural dialogue - „Holy Grail" of AI

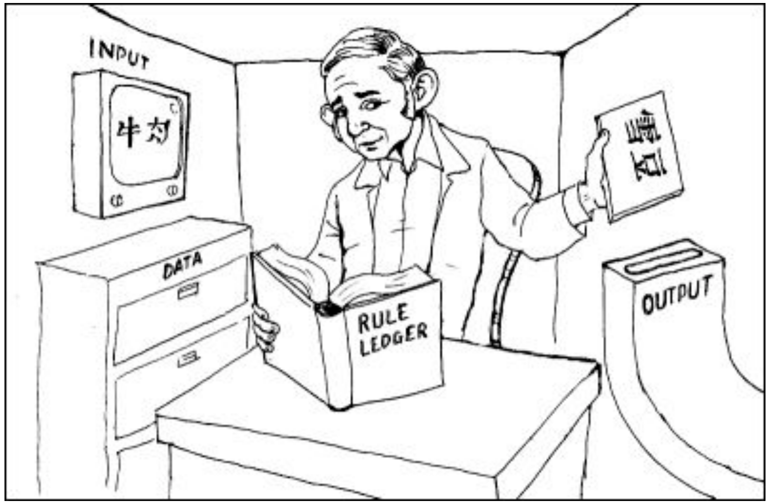
Turing Test



I propose to consider the question "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think."
[Turing, 1950]

Critics: Understanding not really needed (no intelligence?)

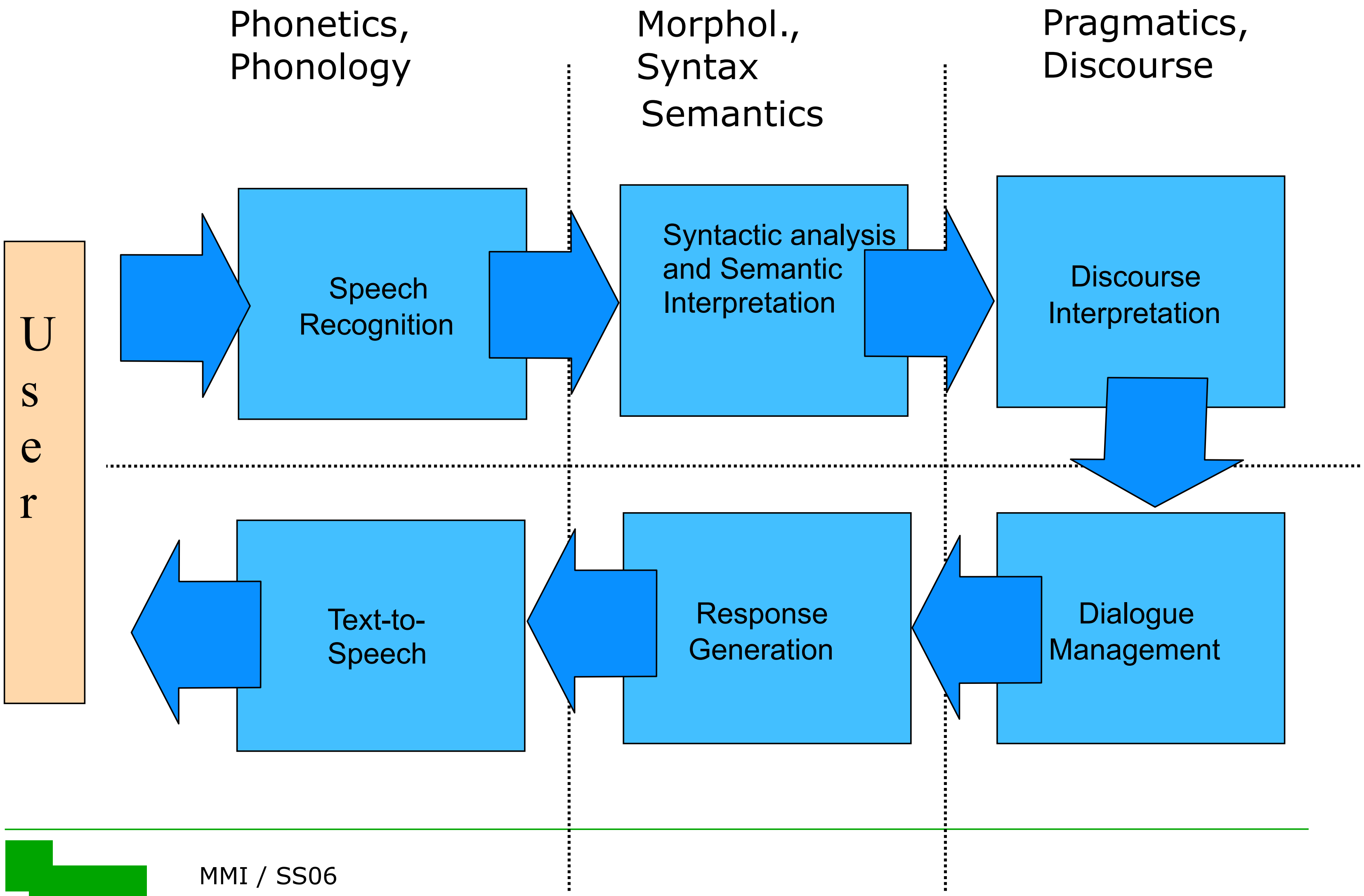
- "Chinese Room" (Searl, 1980)
- ELIZA (Weizenbaum, 1966)



Natural language – levels to look at

- *Phonology and Phonetics*
  study of speech sounds and their usage
- *Morphology*
  study of meaningful components of words
- *Syntax*
  study of structural relationship between words
- *Semantics*
  study of meaning, of words (lexical semantics) and of word combinations (compositional semantics)
- *Pragmatics*
  study of how language is used to accomplish goals (said: „I‘m cold“ → meant: „shut the window“)
- *Discourse*
  study of linguistic units larger than single utterances

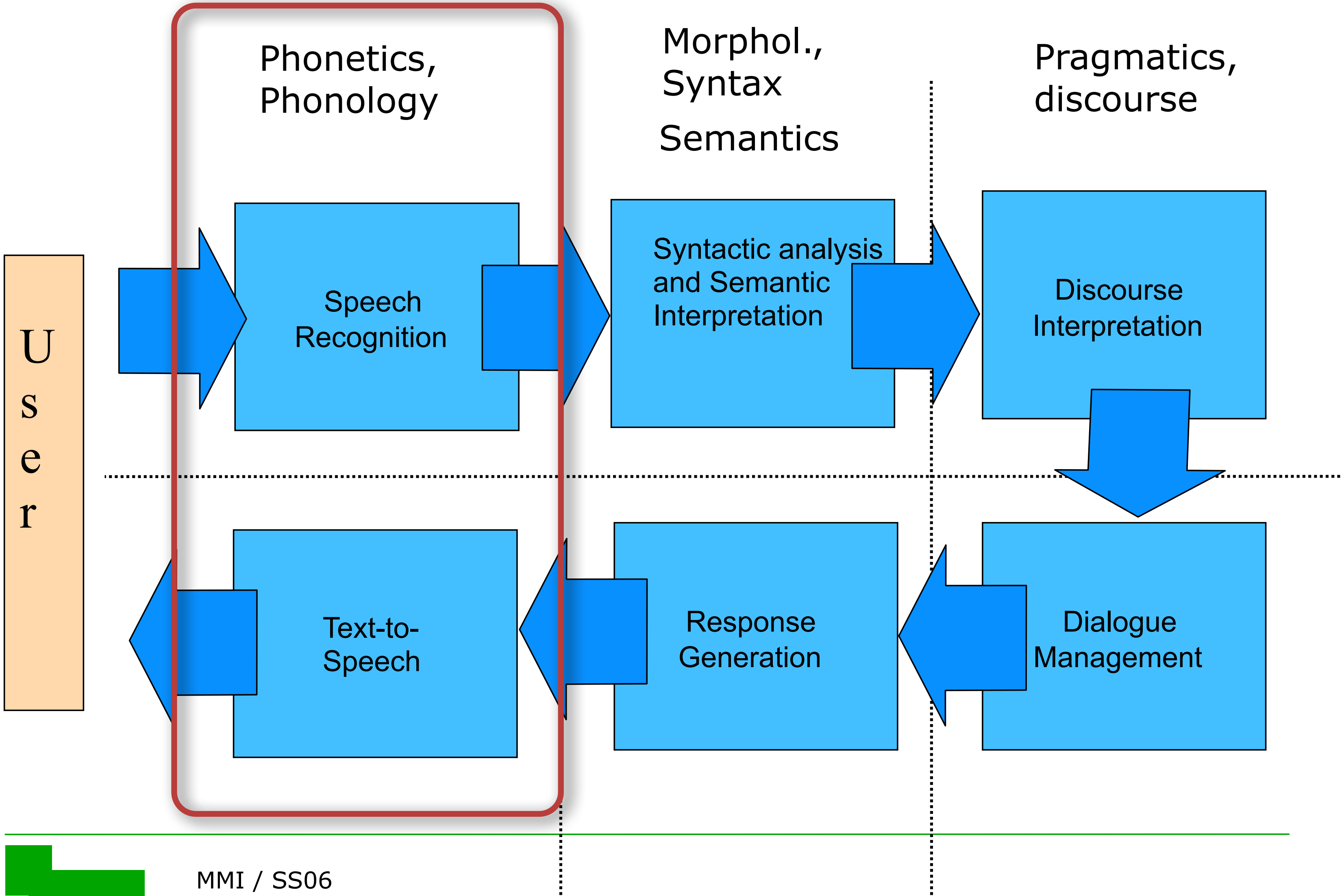
Classical SLDS



Spoken Dialogue System - overview

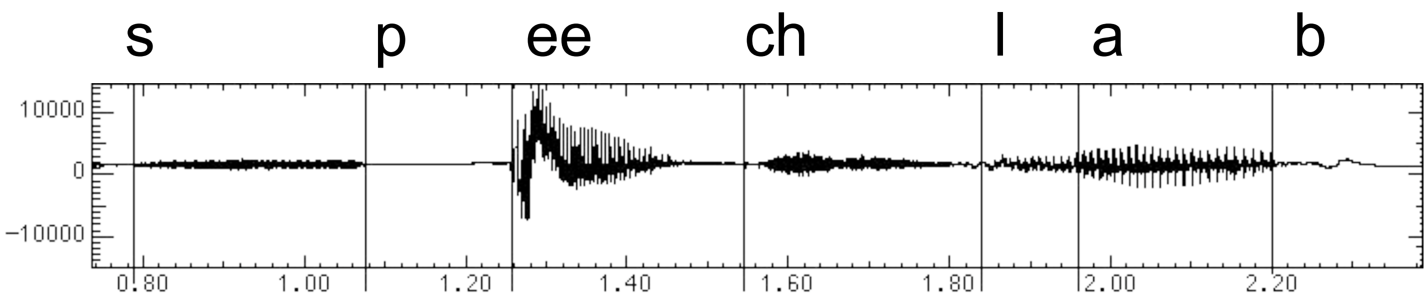
- Speech Recognition:
  - Decode the sequence of feature vectors into a sequence of *words*.
- Syntactic Analysis and Semantic Interpretation:
  - Determine the utterance *structure* and the *meaning* of the words.
- Discourse Interpretation:
  - Understand what the *utterance means* and what the user *intends* by interpreting in *context*.
- Dialogue Management:
  - Determine *goals* and *plans* to be carried out to respond properly to the user intentions.
- Response Generation:
  - Turn communicative act(s) into a *natural utterance*
- Text-to-speech:
  - Turn the words into *synthetic speech*

Spoken Dialogue System



Starting and end point: acoustic waves

- Human speech generates a wave
- A wave for the words “speech lab”:



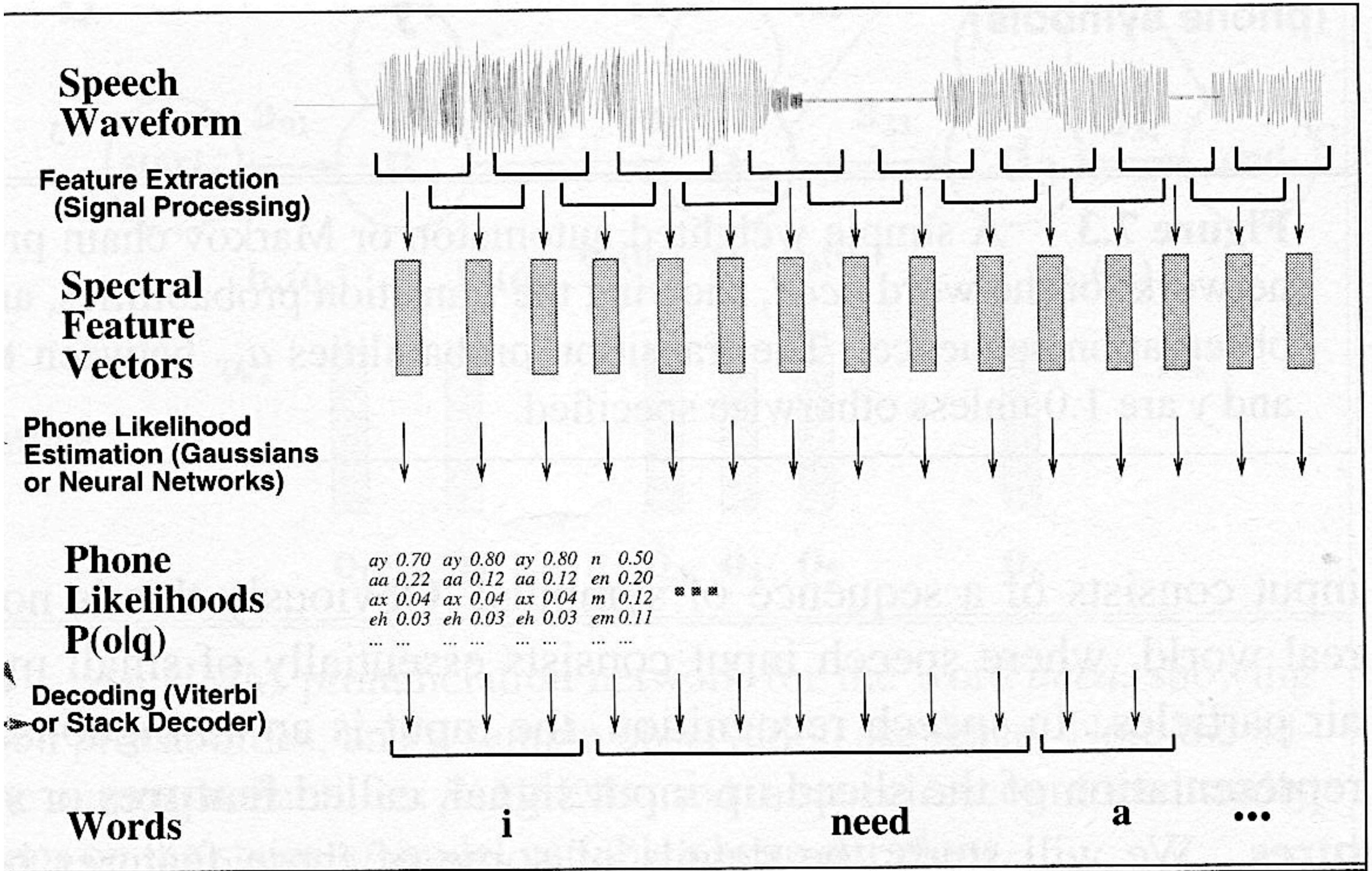
Basics

- *Phonetics*: study of speech sounds
  - *Phone* (*segment*) = speech sound (e.g. „[t]")
  - Phones = *vowels*, *consonants*
  - *Diphone*, *triphone*, … = combination of phones
  - *Syllables* = made up of vowels and consonants, not always clearly definable („syllabification problem")
  - *Prominence* = *Accented* syllables that stand out
    - Louder, longer, pitch movement, or combination
  - *Lexical stress* = accented syllable if word is accented
    - „CONtent" (noun) vs „conTENT" (adjective)
  - *Allophone:* different pronounciations of one phone
    - [t] in „tunafish" → aspirated, voicelessness thereafter
    - [t] in „starfish" → unaspirated

Basics cont.

- *Phonology*: describes the systematic ways that sounds are differently realized
  - *Phoneme* = smallest meaning-distinctive, but *not meaningful* articulatory unit
    - Phones [b] (`bill´) and [ph] (`pill´) discriminate two meanings → different phonemes /b/ und /p/
    - Subsume different elemental sounds under one phoneme, e.g. [p] in `spill´ and [ph] in `pill´ → /p/
  - *Phonological rules* = relation between phoneme and its allophones
  - Every language has ist own set of phonemes and rules

Speech recognition

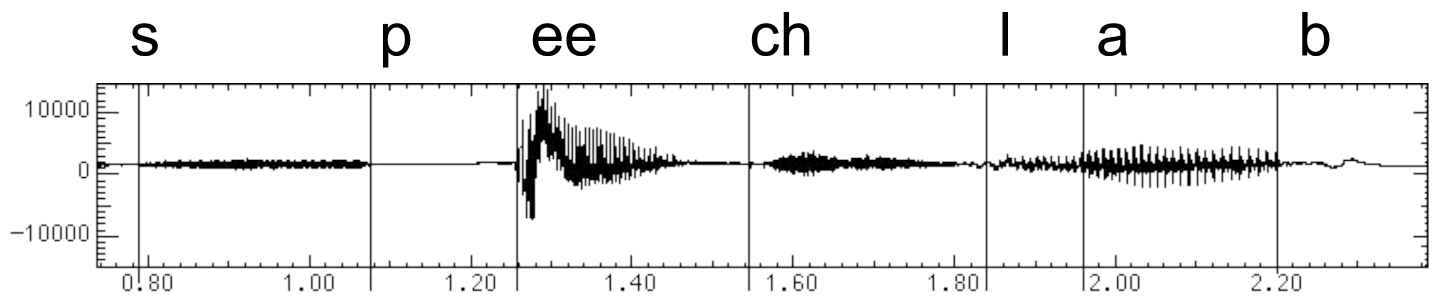


2 Schematic architecture for a (simplified) speech recognizer.

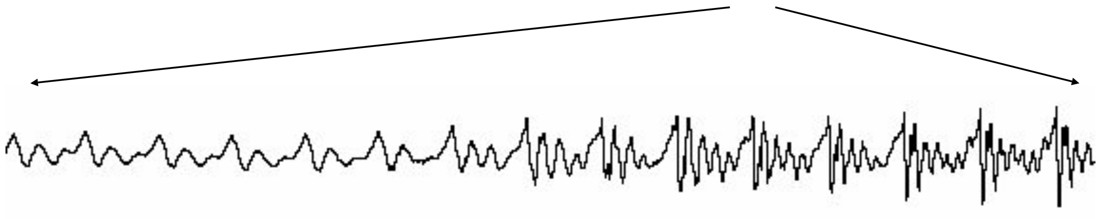
(Jurafsky & Martin, 2000)

Acoustic Waves

- A wave for the words “speech lab” looks like:

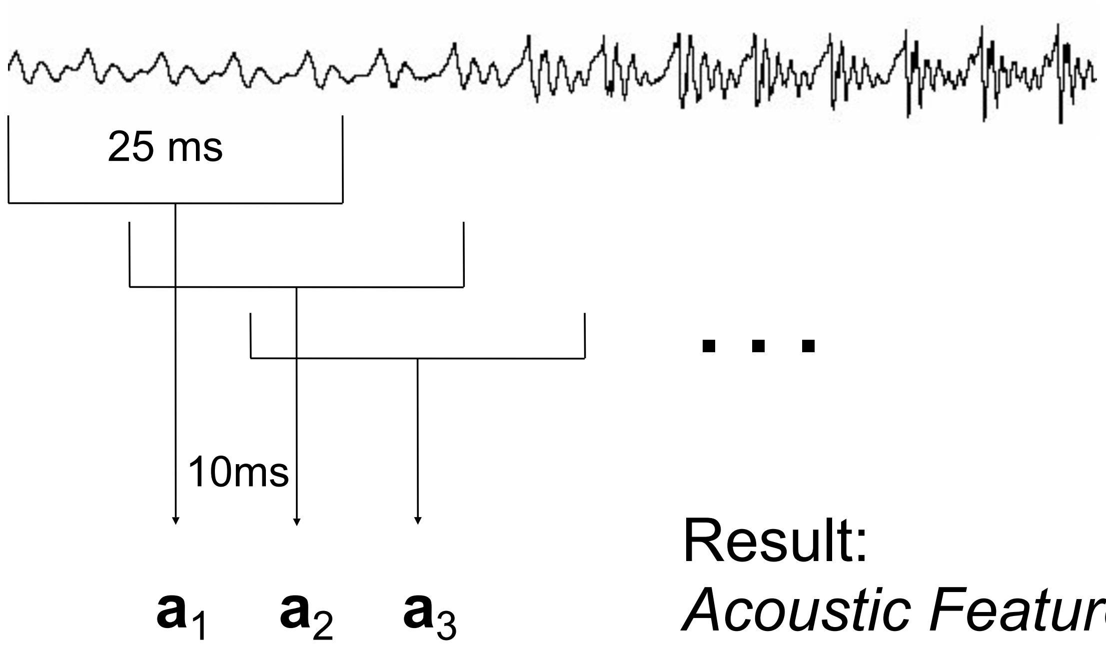


“l” to “a” transition:



Acoustic Sampling

- 10 ms frame (= 1/100 second)
- ~25 ms window around frame to smooth signal processing



Result:
Acoustic Feature Vectors

The Speech Recognition Problem

- Recognition problem
  - Find most likely sequence **w** of "words" given the sequence of acoustic observation vectors **a**

- Use *Bayes' law* to create a generative model
  - $P(a,b) = P(a|b)\, P(b) = P(b|a)\, P(a)$
  - Joint probability of *a* and *b* = a priori probability of *b* times the probability of *a* given *b*

- Apply to recognition problem:
  - *acoustic model*: $P(\mathbf{a}|\mathbf{w})$ (→ HMMs for subword units)
  - *language model*: $P(\mathbf{w})$ (→ Grammars, etc.)
  - $\text{ArgMax}_{\mathbf{w}}\ P(\mathbf{w}|\mathbf{a}) = \text{ArgMax}_{\mathbf{w}}\ P(\mathbf{a}|\mathbf{w})\, P(\mathbf{w}) / P(\mathbf{a})$

    $= \text{ArgMax}_{\mathbf{w}}\ P(\mathbf{a}|\mathbf{w})\, P(\mathbf{w})$

Crucial properties of ASRs

- Speaker:
  - independent vs. dependent
  - adapt to speaker vs. non-adaptive
- Speech:
  - recognition vs. verification
  - continuous vs. discrete (single words)
  - spontaneous vs. read speech
  - large vocabulary (2K-200K) vs. limited (2-200)
- Acoustics
  - noisy environment vs. quiet environment
  - high-res microphone vs. phone vs. cellular
- Performance
  - real time, low vs. high Latency
  - anytime results vs. final results

Text-to-speech

Text-to-speech

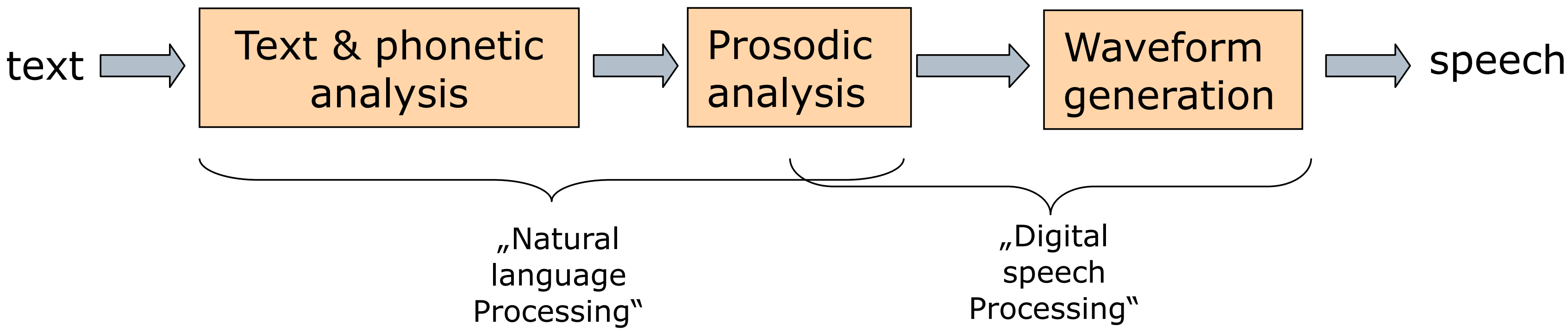
- Mapping text to phones

- The simplest (and most common) solution is to record prompts spoken by a (trained) human
- Produces human quality voice
- Limited by number of prompts that can be recorded
- Can be extended by limited cut-and-paste or template filling

Text-to-speech

Central steps:

1. Analyse text and select sound *segments*
2. Determine prosody and how to model it with single segments
3. Turn into acoustic waveform (*speech synthesis*)



Crucial choice: which segments?

Co-articulation = change in segments due to movement of articulators in neighboring segments

- Phonemens?
  - problematic due to co-articulatory effects
- Allophones
  - Variants of a phoneme in specific contexts
  - Example: Phoneme /p/ → [p] in spill and [ph] in pill
- Diphones („Zweilautverbindungen")
  - Diphones start half-way thru 1st phone and end half-way thru 2nd
  - ⇒ critical phone transition is contained in the segment itself, need not be calculated by synthesizer
  - Example: diphones for German word „Phonetik":
    f-o, o-n, n-e, e-t, t-i, i-k

Phonetic analysis

from words to segments

- Look up pronunciation dictionary
- Words/wordforms
  - e.g. CMUdict: ~125.000 wordforms
  - primary stress, secondary stress, no

  http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Word	Pronunciation
goose	[gus]
geese	[gis]
hedgehog	[ˈhɛdʒ.hɔg]
hedgehogs	[ˈhɛdʒ.hɔgz]

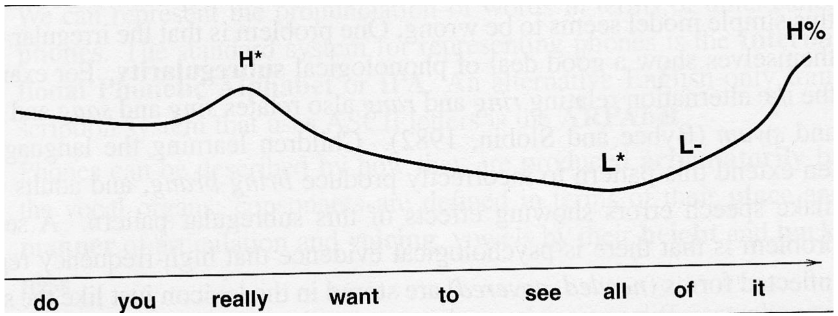
- always a lot of unknown words left
- map letters to sounds with rules
  - MITalk (1987): 10.000 rules repository: p – [p]; ph – [f]; phe – [fi]; phes – [fiz]; … … …
  - Festival: rules account for co-articulation: [c h] + any consonant = `k´, else `ch´ (`christmas´ vs. `choice´)
  - Usually machine learned from large data sets

Prosodic analysis

from words+segments to boundaries, accent, F0, duration

TTS systems need to create proper prosody by adapting:

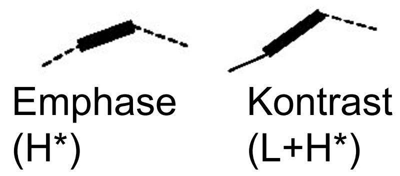
- Prosodic phrasing/boundaries:
  - Break utterances into units
  - Punctuation and syntactic structure useful, but not sufficient
- Duration of segments:
  - Predict duration of each segment
  - Helps to create prominence
- Intonation/accents on/over segments:
  - Predict accents: which syllables should be accented?
  - Realize as F0 contour („pitch") with special form for accents



Pitch *accents*

- In the first place, properties of *words*
- Decisive for how words are interpreted, used to…
  - emphasize new information (“Then I saw a **church**.”)
  - contrast parts („I like **blue** tiles better than **green** tiles.“)
  - explicitly focus parts („I said I saw a **church**.“)

- Different pitch accents serve different functions in discourse



- Which to choose depends on content and context
  - Given (topic, theme) or new information (rheme)?
  - Information mutually agreed or not?
  - → „concept-to-speech, content-to-speech“

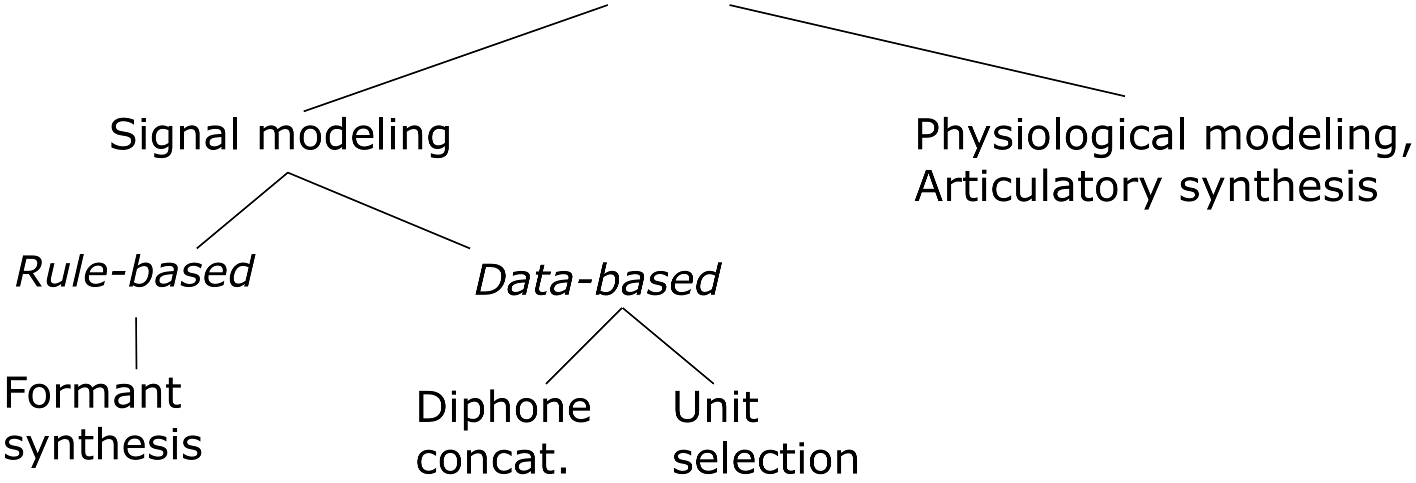
Duration

Generate segments with appropriate duration. Influenced by

- Segmental identity
  - /ai/ in ‘like’ twice as long as /I/ in ‘lick’
- Surrounding segments
  - vowels longer following voiced fricatives than voiceless stops
- Syllable stress
  - stressed syllables longer than unstressed
- Word “importance”
  - word accent with major pitch movement lengthens
- Location of syllable in word
  - word ending longer than starting longer than word internal
- Location of the syllable in the phrase
  - phrase final syllables longer than in other positions

Waveform synthesis

from segments, f0, duration to waveform



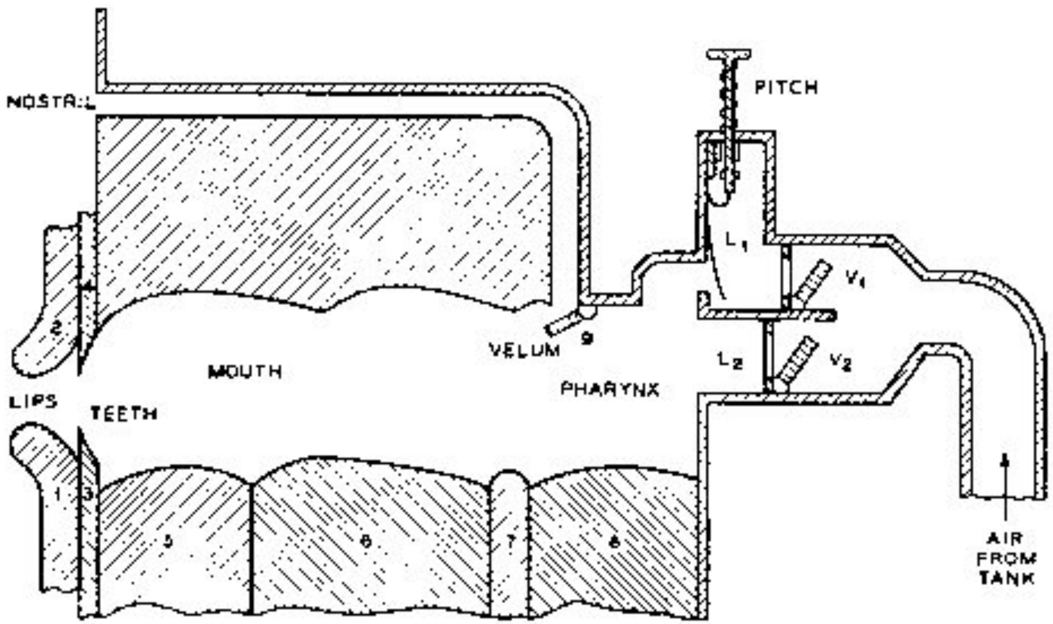
Start with acoustics, rules to create formants

Use databases of stored speech to assemble new utterances

Model movements of articulators and acoustics of the human vocal tract

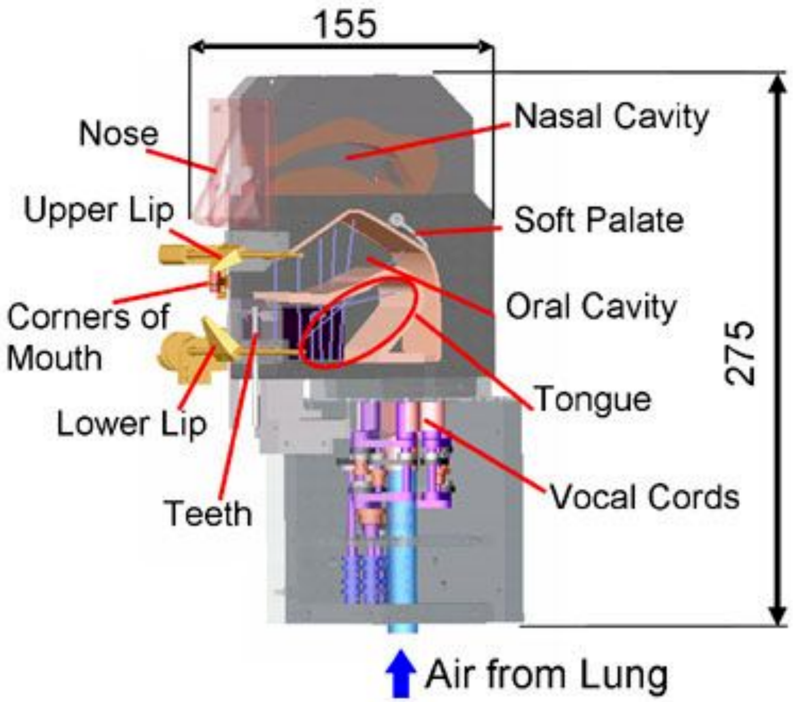
Articulatory synthesis

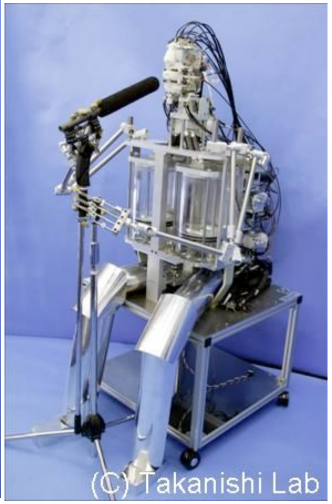
- based on physical or nowadays computational models of the human vocal tract and the articulation processes occurring there
- few of them currently sufficiently advanced or computationally efficient



Articulatory synthesis

Talking robots WT-4, WT-5
Waseda University, Tokyo

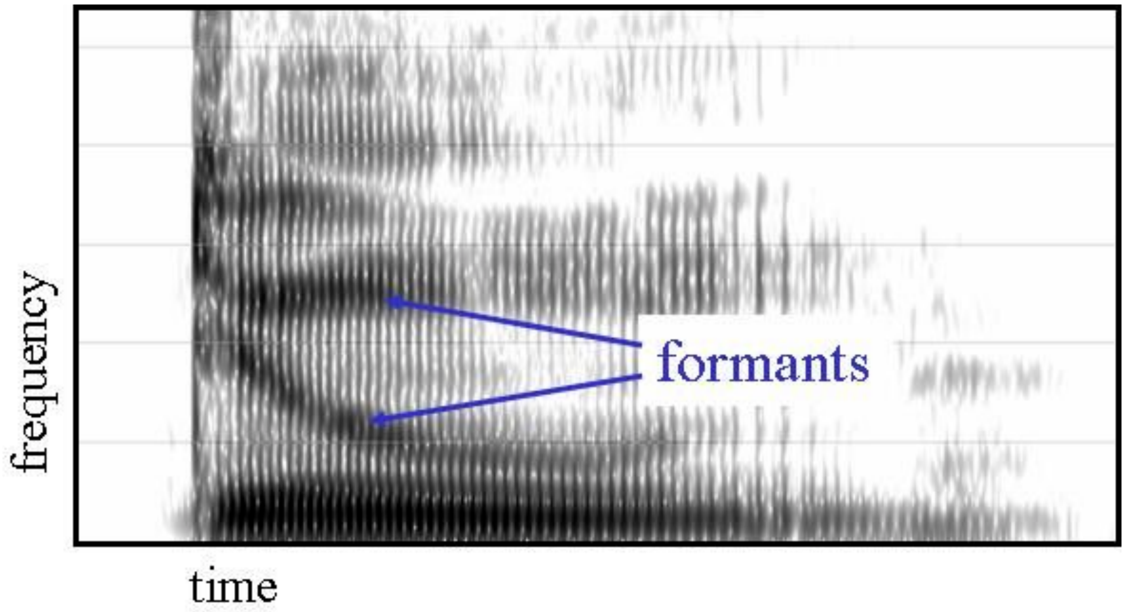




„sasisuseso“

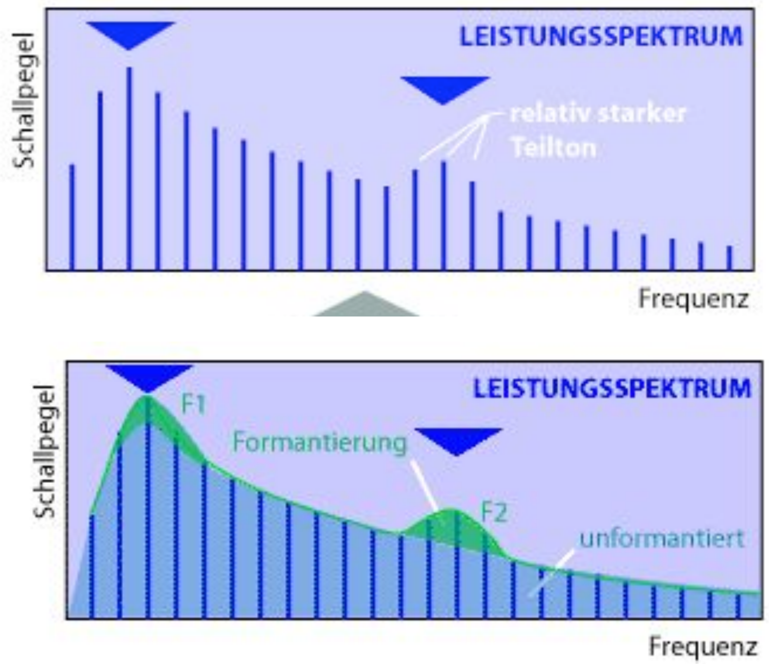
Formant synthesis

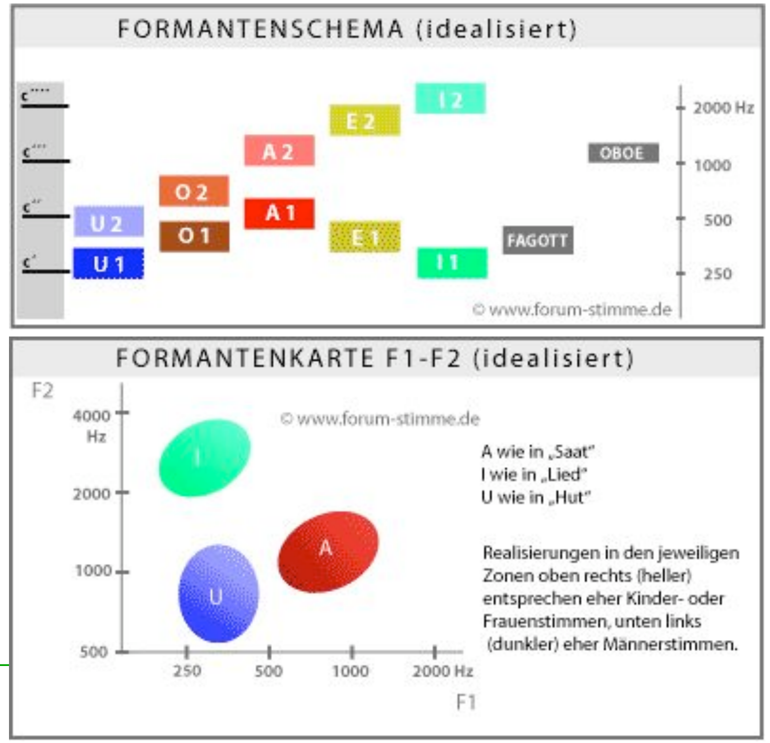
- *Formant*: Frequenzregion, in der die dort hineinfallenden Teiltöne besonders stark sind
- Wesentlichen Elemente der Klangbildung, je nach Lage und Stärke verschiedene Vokale und Timbre



Formant Synthesis

- Annahme: Die für die menschliche Perzeption wesentliche Information ist durch die Töne in den Formanten kodiert
- Dabei prägen vor allem die beiden am tiefsten gelegenen Formanten (F1, F2) die Lautwahrnehmung, mitunter reicht zur Wahrnehmung bestimmter Vokale auch nur ein Hauptformant

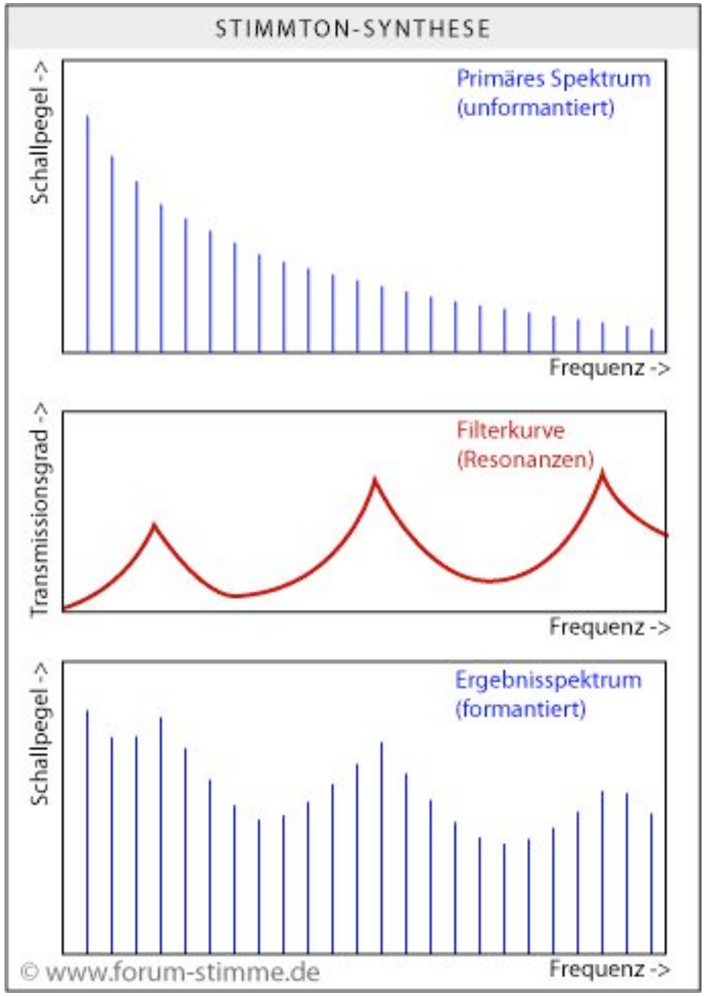




www.forum-stimme.de

Formant Synthesis

- Rules model relations between tones and acoustic features
- Advantages
  - flexibilty
  - not much storage space needed
- Disadvantages
  - Sounds mechanical
  - Complicated rule sets
- Most common systems while computers were relatively underpowered
  - 1979 MIT MITalk (Allen, Hunnicut, Klatt),
  - 1983 DECtalk system, 'Klatt synthesizer'



Data-based synthesis

- Nowadays all current commercial systems (1990's-)
- Steps:
  1. Record basic inventory of sounds (offline)
  2. Retrieve sequence of units at run time (at run-time)
  3. Concatenate and adjust prosody (at run-time)

- What kind of units?
  - Minimize context contamination, capture *co-articulation*
  - Enable efficient search
  - Segmentation and concatenation problems
- How to join the units?
  - dumb (just stick them together)
  - PSOLA (Pitch-Synchronous Overlap and Add), MBROLA (Multi-band overlap and add)

Einheitenlänge	Einheit	#Einheiten (Englisch)	#Regeln	Qualität
kurz	Allophone	60-80	hoch	gering
	Diphone	$<40^2-65^2$		
	Triphone	$<40^3-65^3$		
	Halbsilben	2K		
	Silben	11K		
	Doppelsilben	<11K2		
	Wort	100K-1.5M		
	Phrasen	∞		
lang	Satz	∞	gering	hoch

Source: E. Andre

Diphone synthesis

- Units = diphones
  - Phones are more stable in middle than at the edges
- Typically 1500-2000 diphones, reduce number
  - *phonotactic constraints*: constraints on the way in which phonemes can be arranged to form syllables
  - collapse in cases of no co-articulation
- Record 1 speaker saying each diphone
  - “Normalized”: monotonous, no emotions, constant volume

- Example: MBROLA (Dutoit & Leich, 1993) http://tcts.fpms.ac.be/synthesis/mbrola.html

Example: TTS for *Max*

- TXT2PHO (IKP) → lexical stress, neutral prosody
- MBROLA + German diphon database
- SABLE tags for additional intonation commands

Phonetic text:

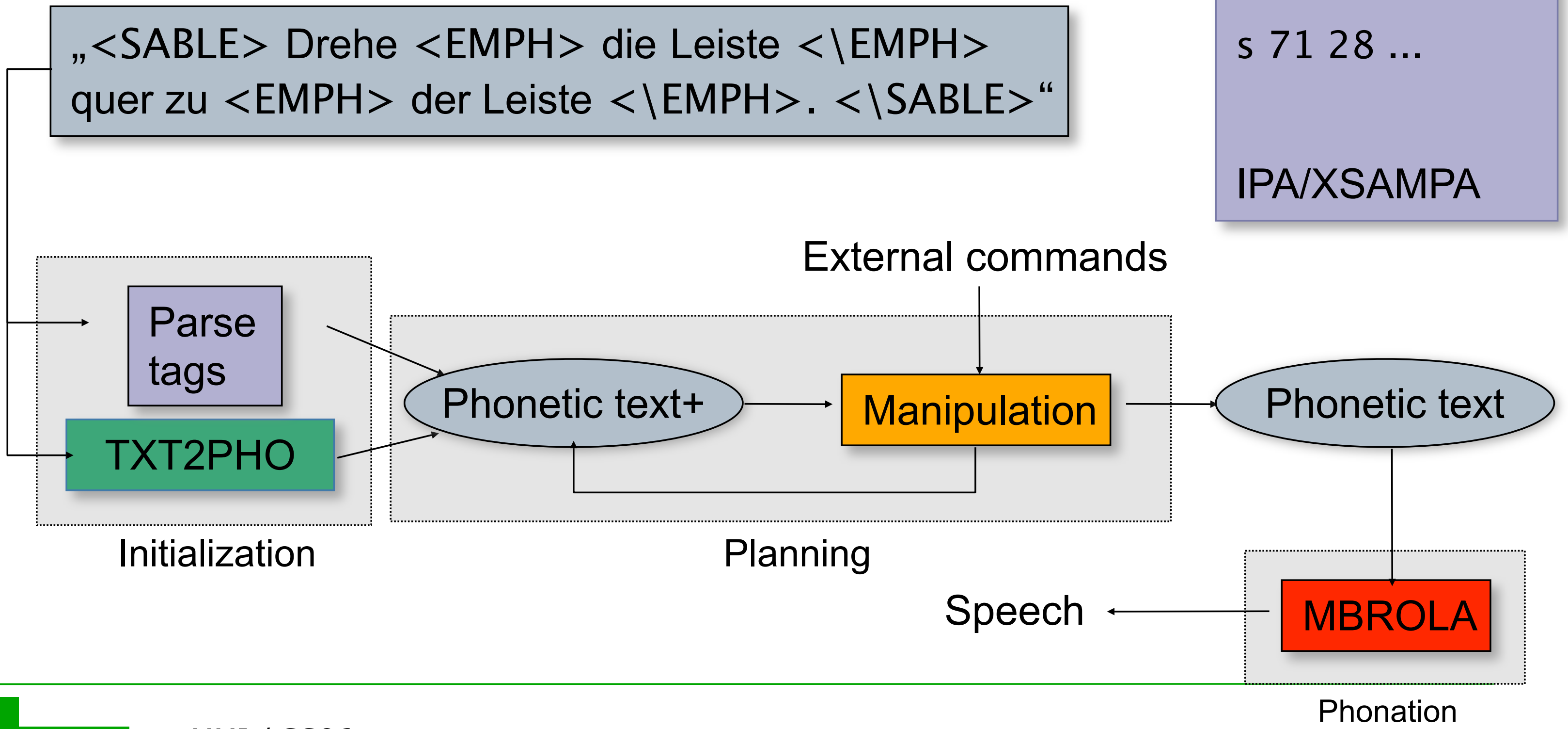
S 105 18 ...

P 90 8 153

a: 104 4 ...

s 71 28 ...

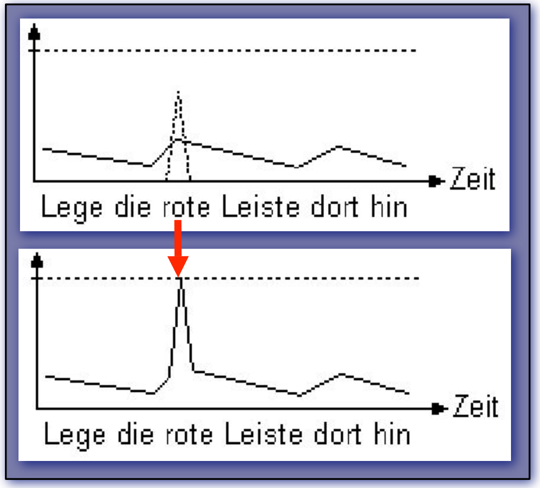
IPA/XSAMPA

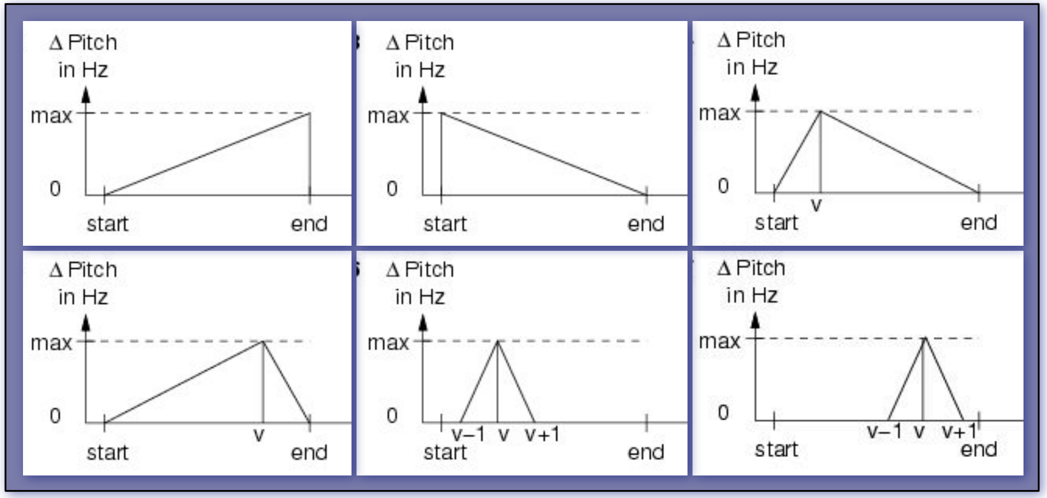


Example: TTS for *Max*

Manipulation of phonetic text

- Overlay stereotyped contours to create accents + durations
- No suprasegmental analysis
- Flexible form, height, duration





Beispiel: Kontrastierung

Wer arbeitet in Bielefeld?

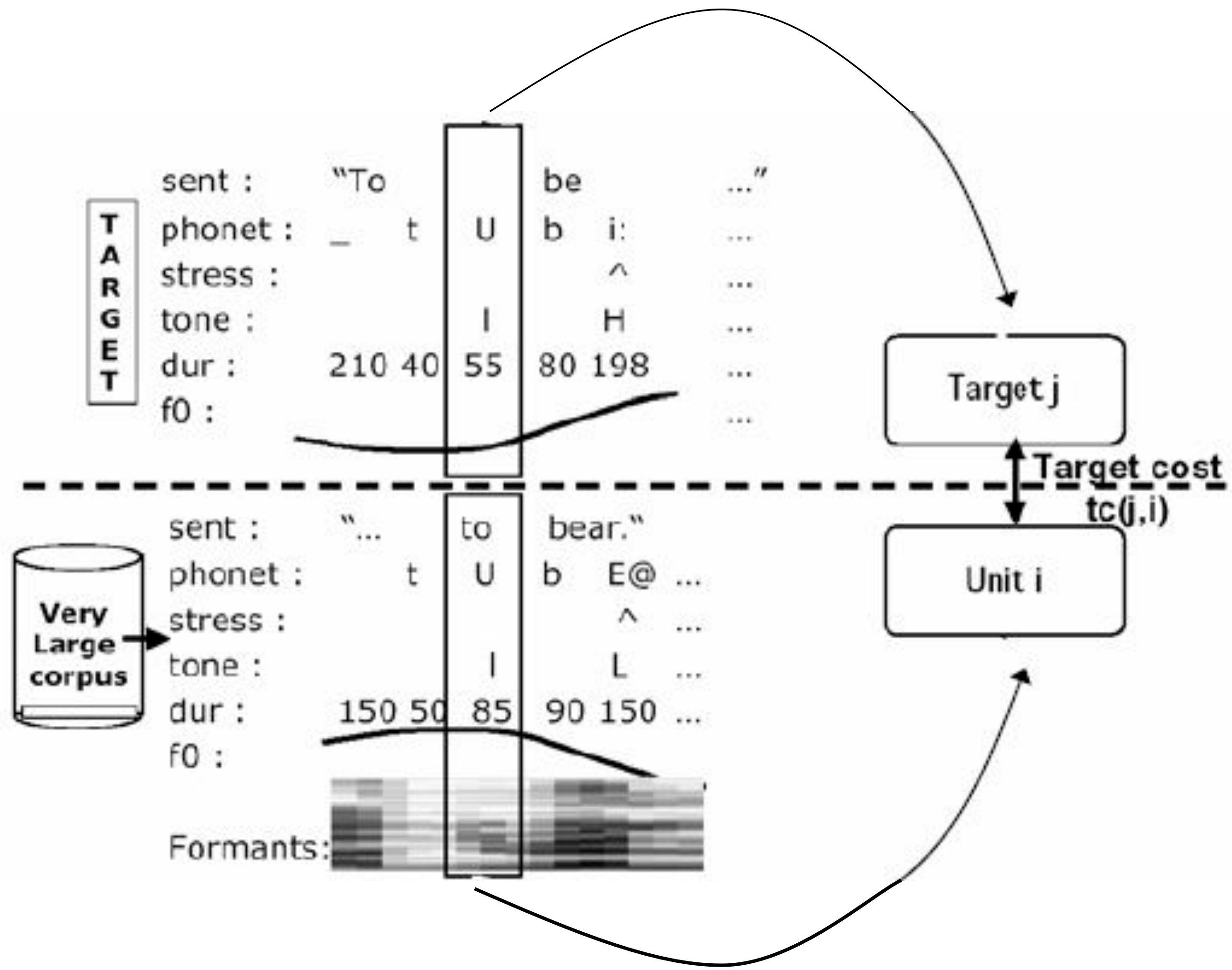
Wo arbeitest du?

Was tust du in Bielefeld?

Unit selection

- One example of a diphone is not enough!
- Unit selection:
  - Record multiple copies of each unit with different pitches and durations
  - How to pick the right units? Search!
  - Example (Hunt & Black, 1996):
    - Input: three F0 values per phone
    - Database: phones+duration+3 pitch values
    - Cost-based selection algorithm
- Non-uniform unit selection
  - Units of *variable* length
  - Reduced need of automatic prosody modeling

Unit selection



Academic TTS systems - demos

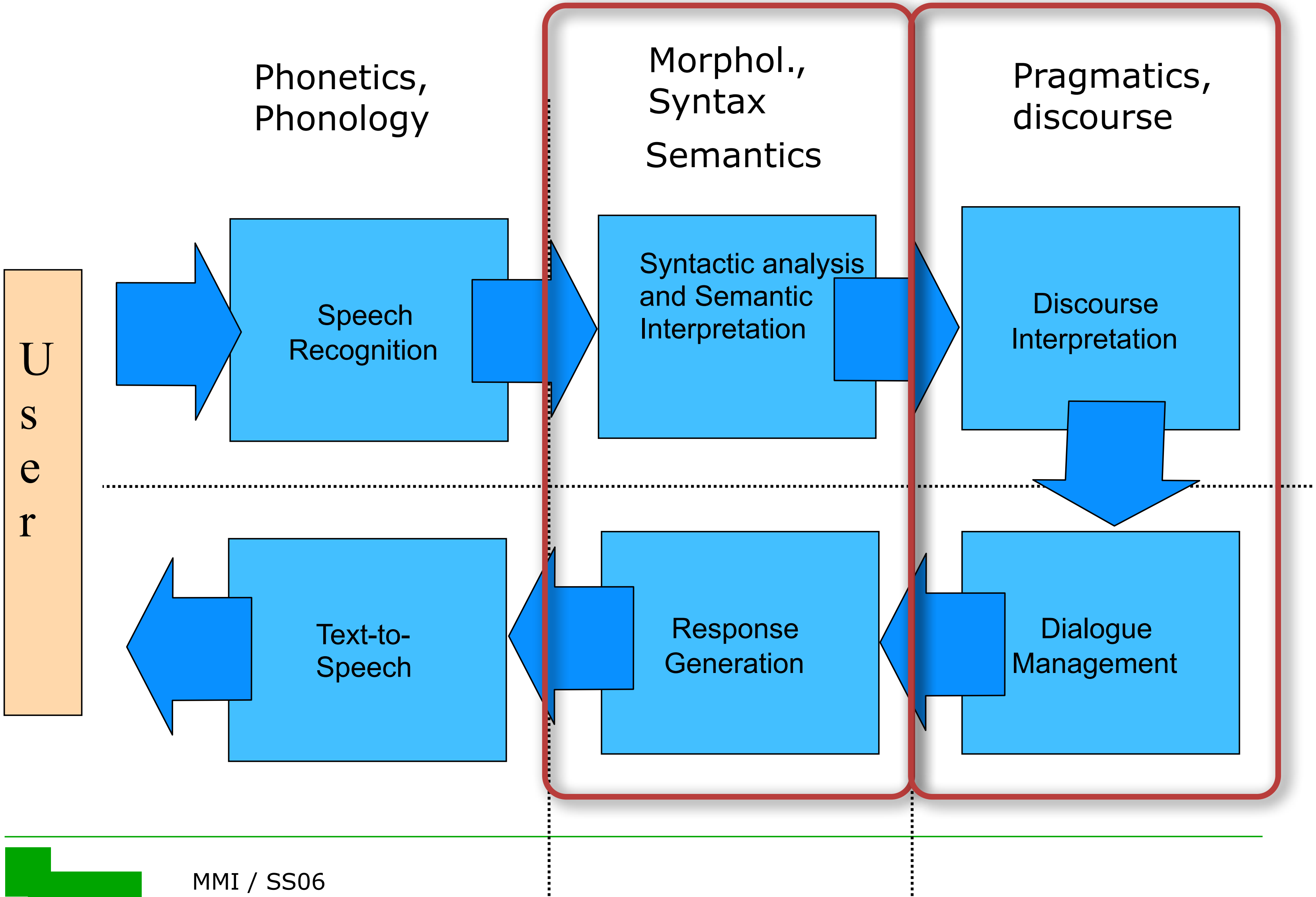
BOSS (IKP, Bonn)	non-uniform unit-selection	Mp3 (2001)
IMS Stuttgart	Diphone concat., Festival+MBROLA	Mp3 (2000)
Uni Duisburg	Formant synthesis	Mp3 (1996)
Mary (DFKI)	Diphone synthesis, MBROLA	Mp3 (2000)
VieCtoS (ÖFAI, Wien)	Halbsilben, schlechte Tobi-Labelung	Mp3 (1998)
SVox (ETH Zürich)	Diphone concat.,	Mp3 (1998)
HADIFIX (IKP, Bonn)	HSlbsilben, DIphone und sufFIXe	Mp3 (1995)

Commercial TTS systems - demos

BabelTech Babil	Diphone concat., MBROLA-like	Mp3 (2000)
AT&T	non-uniform unit-selection	Mp3 (1998)
BabelTech BrightSpeech	non-uniform unit-selection	Mp3 (2003)
IBM ctts	non-uniform unit-selection	Mp3 (2002)
Loquendo	non-uniform unit-selection	Mp3 (2003)
Nuance RealSpeak	non-uniform unit-selection	Mp3 (2006)
SVox Corporate	Diphone concat.	Mp3 (2005)

- Comparison of state-of-the-art TTS systems
  http://ttssamples.syntheticspeech.de/deutsch/index.html
- Janet Cahn‵s Master Thesis, PhD Thesis
  http://xenia.media.mit.edu/~cahn/
- Demos and links for speech synthesizers
  http://felix.syntheticspeech.de/
- Lecture on speech synthesis by Bernd Möbius http://www.ims.uni-stuttgart.de/~moebius/teaching.shtml

Next week:



Phonetics,
Phonology
Morphol.,
Syntax
Semantics
Pragmatics,
discourse
U
s
e
r
Speech
Recognition
Syntactic analysis
and Semantic
Interpretation
Discourse
Interpretation
Text-to-
Speech
Response
Generation
Dialogue
Management

Student projects

- **Aufgabe 1:** Evaluiere die Usability eines Systems/Interface nach einem der drei Standardverfahren. Die Evaluationen soll jeweils geplant, durchgeführt und ausgewertet werden. Alle Schritte sind schriftlich zu dokumentieren (max. 10 Seiten) und bis zum 21.6. abzugeben.

- **Aufgabe 2**: Verwende die Ergebnisse Deiner Evaluation, um einen Vorschlag für ein verbessertes Design zu machen. Skizzieren dazu einen Papierprototypen (Vorschlag zur Interface- und Interaktionsgestaltung auf Papier). Abgabe bis 30.6.

- **Schriftlicher Abschlusstest:**
  - vorauss. Termin: 3.8.06
  - Dauer: 1,5 Std.
  - Anmeldung per EMail an skopp@techfak bis 14.7.06