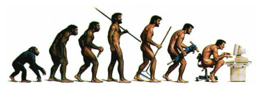
Human-Computer Interaction

Session 8

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

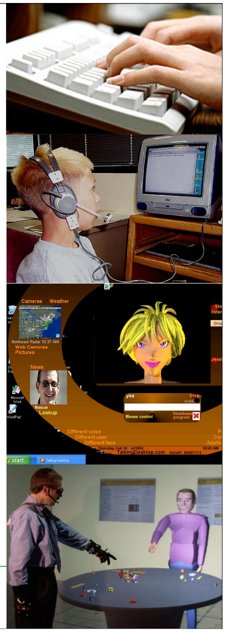
Overview: machines as...

tools ➜ operate

smart tools ➜ instruct

interactive interlocutors ➜ converse

companions ➜ collaborate



Overview: machines as...

tools ➜ operate

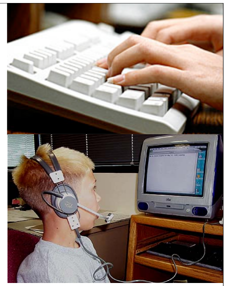
The computer as a multi-functional tool that helps solving problems and achieving tasks

User iteratively operates the computer

- user operates the machine
- machine performs local problem-solving task
- machine gives feedback
- (goto 1)

Build usable tools ➜ user-centered design

Overview: machines as...



tools ➜ operate

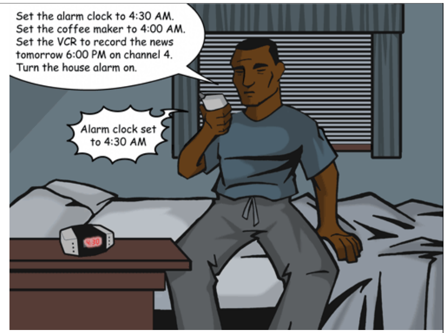
smart tools ➜ instruct

Make tools smarter and more autonomous, such that the user can instruct them most efficiently and easily

Task intelligence ➔ can instruct on abstract levels

Communication intelligence ➔ can instruct naturally

Using speech to interact with systems



Intuitive form of communication, no need for training
- Speech in/speech out

Relates to (one) way of thinking; *but* images, maps, ...
- When possible, ofte combined with other media

Credo: Computer should adopt human way of interaction

Speech interaction

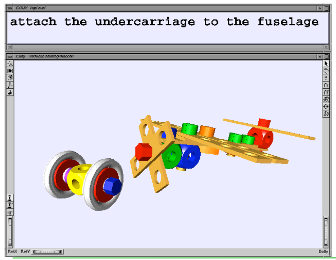
Used today...
- on the desktop, e.g. dictate
- on the phone, e.g. ticket booking, pizza ordering

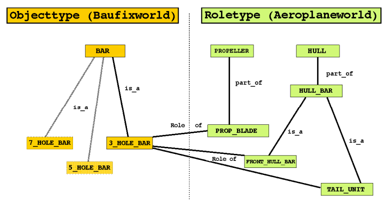
Ongoing research on...
- natural language
- mobile devices & robots
- automotive interaction
- Virtual Reality

Example: Virtual Constructor

(Jung et al., 1998)

Interpret instructions based on knowledge about the world and the situation
- Objects can be refered to by their actual or potential role („tail unit" instead of „bar"), as well as their context-dependent properties





Example: Talking Desktop

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

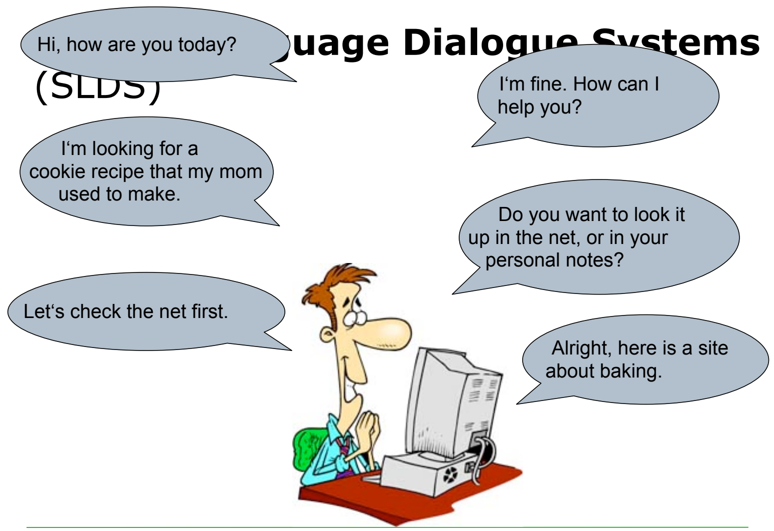
Example: Talking Desktop

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 in a "relatively natural manner"

…uage Dialogue Systems (SLDS)



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

Natural language – things to think of

Phonology & Phonetics
: speech sounds and their usage

Morphology
: components and structure of words

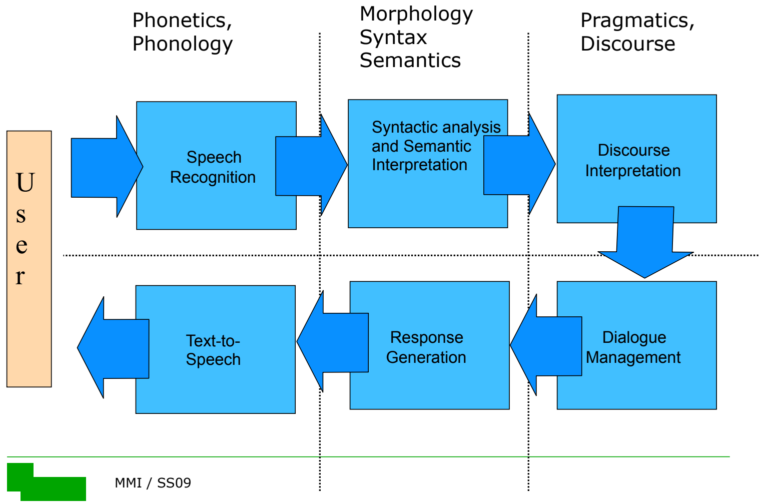
Syntax
: structural relationship between words & phrases

Semantics
: meaning of words (lexical) and word combinations (compositional)

Pragmatics
: language use in context in order to accomplish things (said: „I'm cold" → meant: „shut the window")

Discourse
: larger meaningful connection across linguistic units

Classical structure of SLDS



Classical structure of SLDS

Speech Recognition
: Decode the sequence of feature vectors into a sequence of *words*.

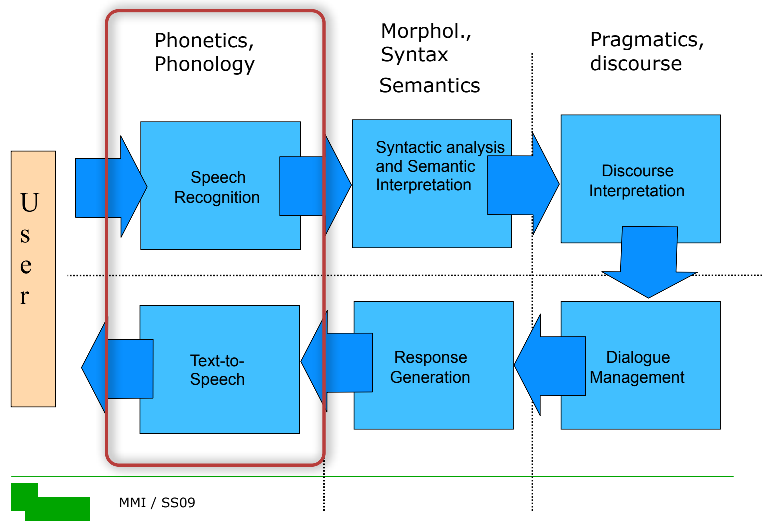
Syntactic Analysis and Semantic Interpretation
: Determine the utterance *structure* and the *meaning* of the words.

Discourse Interpretation
: Understand what the *utterance means* and what the user *intends* by putting it in *context*.

Dialogue Management
: Determine *how to respond* properly to the user intentions.

Response Generation
: Turn communicative act(s) into a *natural utterance.*

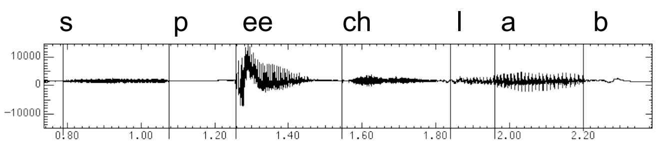
Text-to-speech
: Turn the words into *synthetic speech.*



Starting and end point: acoustic waves

Human speech generates a wave

A wave for the words "speech lab":



Phonetics

study of speech sounds

- *Phone* (*segment*) = speech sound (e.g. „[t]")
  - *vowels*, *consonants*
- *Allophone:* different pronounciations of a phone
  - [t] in „tunafish" → aspirated, voicelessness thereafter
  - [t] in „starfish" → unaspirated
- *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)

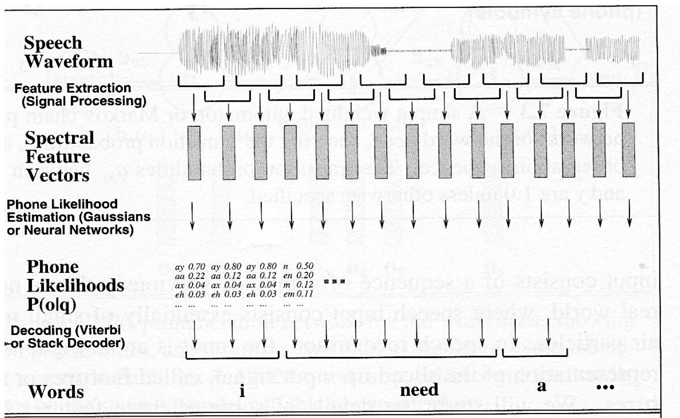
Phonology

study of the ways that sounds are differently realized to make meaning

- *Phoneme* = smallest meaning-distinguishing, 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
  - ~40 German phonemes: /p/, /t/, /k/ (plosives); /m/, /n/, /ŋ/ (nasals); /a:/, /a/, /e:/, /ɛ/ (vowels)

Speech recognition

(in a nutshell)

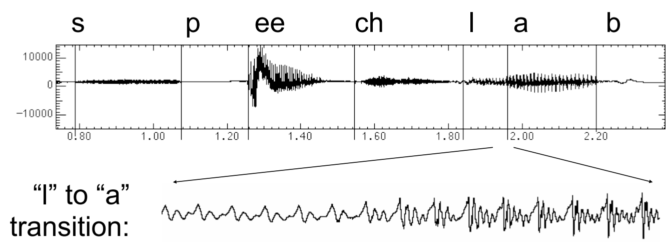


Schematic architecture for a (simplified) speech recognizer.

(Jurafsky & Martin, 2000)

Waveform

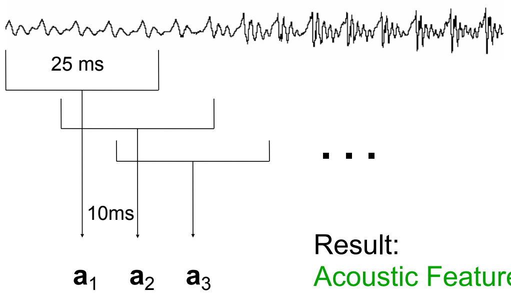
A wave for the words "speech lab" looks like:



Acoustic Sampling

10 ms frame (= 1/100 second)

~25 ms window around frame to smooth signal processing



Result:
Acoustic Feature Vectors

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

Text-to-speech synthesis

Problem: mapping text to audible phones

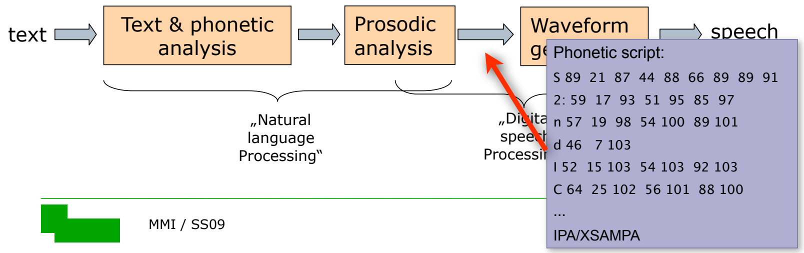
Simplest (and most common) solution

- 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 synthesis

Central steps:

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



Which segments?

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

Phonemes?

- 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

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

Look up words/wordforms in a pronunciation dictionary

- e.g. CMUdict: ~125.000 wordforms
- + primary stress, secondary stress
  http://www.speech.cs.cmu.edu/cgi-bin/cmudict

always a lot of unknown words left: letter-to-sound 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

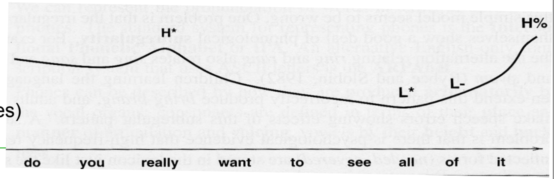
Prosodic analysis

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

TTS systems need to create proper prosody by adapting:

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

ToBI Notation
(Tones and Break Indices)



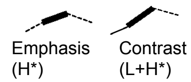
Pitch accents

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



What 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"

Waveform synthesis

from segments, f0, duration to waveform

Signal modeling
- **Rule-based**
  - Formant synthesis
- **Data-based**
  - Diphone concat.
  - Unit selection
  - HMM-based

Physiological modeling, **Articulatory synthesis**

Start with acoustics, rules to create formants

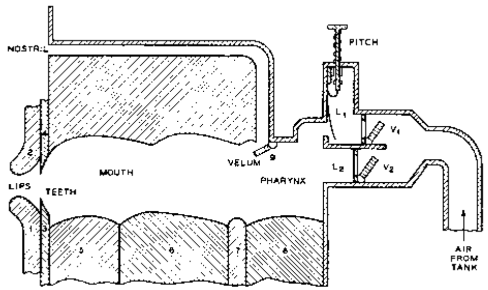
Use databases of stored speech to assemble new utterances

Model movements of articulators and acoustics of the human vocal tract

Articulatory synthesis

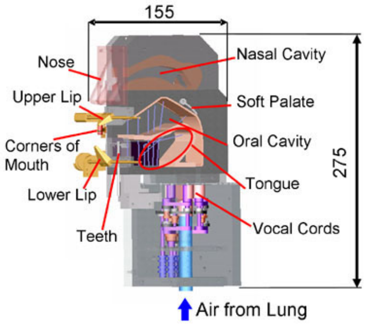
based on physical or (nowadays) computational models of the human vocal tract and the articulation processes occurring there

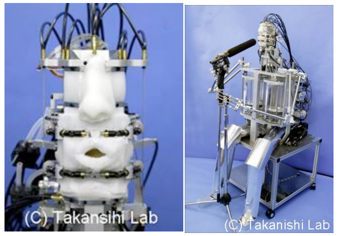
few of them sufficiently advanced or computationally efficient

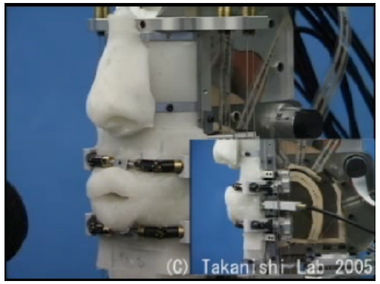


Articulatory synthesis

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





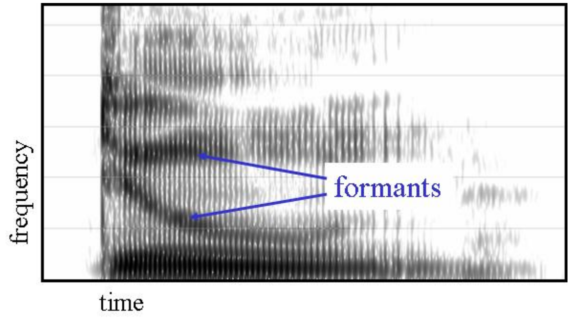


„sasisuseso"

Formant synthesis

Formant = Region of frequency in which tones have a (comparably) strong intensity

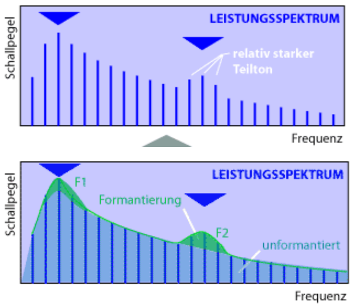
Significant elements of tone, depending on position and intensity of the vowel and timbre

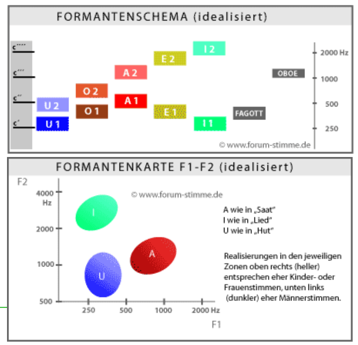


Formant Synthesis

Assumption: Important perceptual information encoded in formants

In particular the first two formants (F1, F2) determine speech perception; sometimes the primary formant is sufficient by itself





www.forum-stimme.de

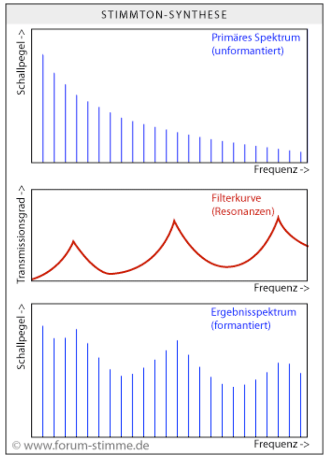
Formant Synthesis

Rules model relations between tones and acoustic features

Advantages
- flexibilty
- not much storage space needed

Disadvantages
- Sounds mechanical
- Complicated rule sets

Common while computers were relatively under-powered
- 1979 MIT MITalk (Allen, Hunnicut, Klatt),
- 1983 DECtalk system, 'Klatt synthesizer'



Data-based synthesis

Almost all current commercial systems (1990's-)

Steps:

1. Record basic inventory of sounds (offline)
2. Retrieve sequence of units at run time (run-time)
3. Concatenate and adjust prosody (run-time)

What kind of units?

- Minimize context contamination, but 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)

Einheiten-länge	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, need to 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 one speaker saying each diphone

- "Normalized": monotonous, no emotions, constant volume

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

Unit selection

One example of a diphone is not enough!

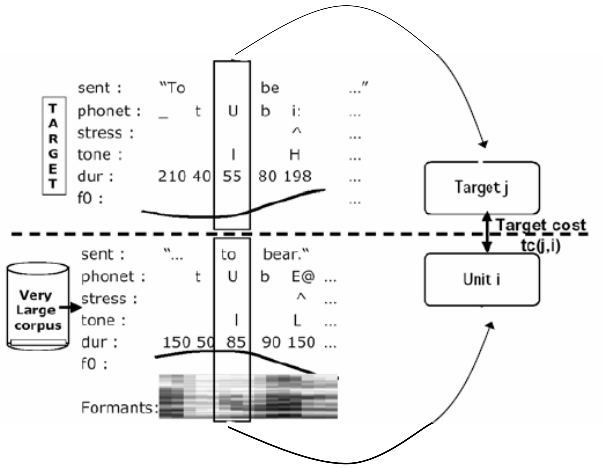
Unit selection:

- 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



HMM-based synthesis

From a sequence of phonemes (+contextual annotation), use HMMs to generate sequences of a parameterised form, from which a waveform can generated

Parameterised form contains information about

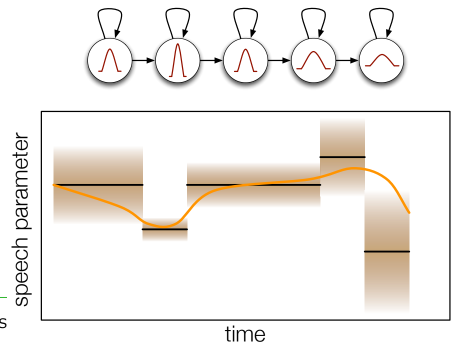
- spectral envelope
- fundamental frequency (F0)
- aperiodic (noise-like) components (e.g. for 'sh' and 'f')

Trajectory HMM algorithm (Tokuda et al.): uses statistics of the dynamic properties during the generation process (instead of generating means of Gaussian)

HMM-based synthesis

Generate most likely observation sequence from HMM

- take statistics of not only the static coefficients, but also the delta and delta-delta too
- Maximum Likelihood Parameter Generation Algorithm



HMM-based synthesis

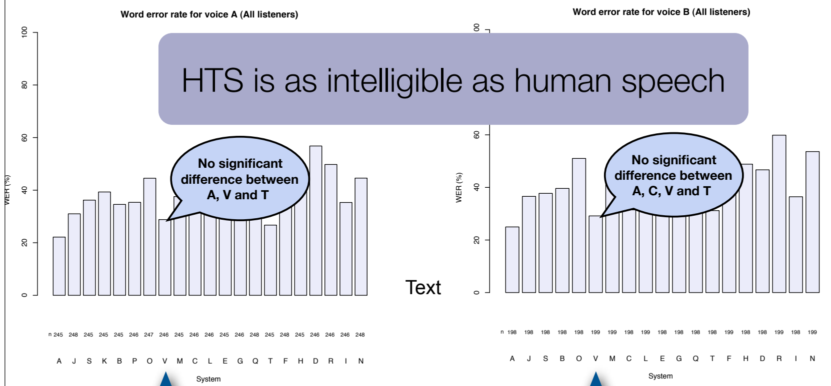
„Full context models": phonetic and prosodic factors

One 5-state HMM for each phoneme, in every required context ➔ sparsity problem!

To synthesise a given sentence:

1. predict linguistic specification (phonetic & prosodic analyses)
2. concatenate corresponding phoneme HMMs
3. generate from the resulting overall HMM

HTS is as intelligible as human speech



A natural speech
B Festival benchmark
C HTS 2005 benchmark
V HTS 2008 (aka HTS 2007')

Centre for Speech Technology Research
University of Edinburgh, UK
www.cstr.ed.ac.uk

Academic TTS systems - demos

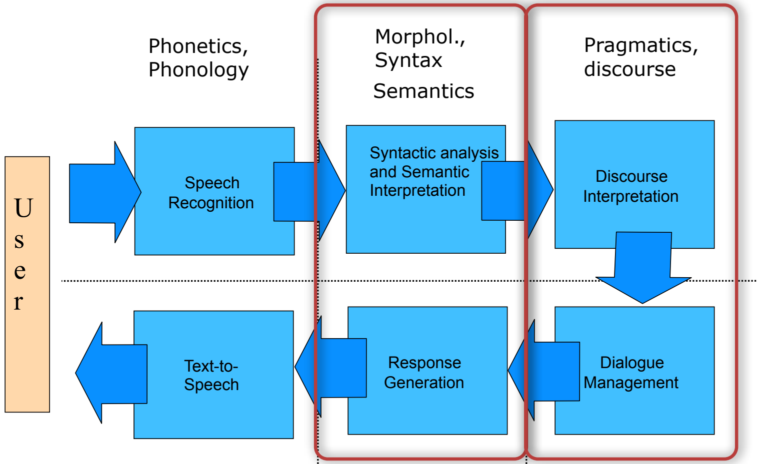
BOSS (IKP, Bonn)	non-uniform unit-selection	Mp3 (2001)
IMS Stuttgart	Diphone concat., Festival+MBROLA	Mp3 (2000)
Uni Duisburg	Formant synthesis	Mp3 (1996)
Mary (DFKI)	Diphone synthesis, HMM	Mp3 (2000) Mp3 (2008)
VieCtoS (ÖFAI, Wien)	Halbsilben, schlechte Tobi-Labelung	Mp3 (1998)
SVox (ETH Zürich)	Diphone 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)

- Bernd Frötschl (FU Berlin): Samples, Tools, Resourcen
  http://page.mi.fu-berlin.de/froetsch/tts.html
- 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