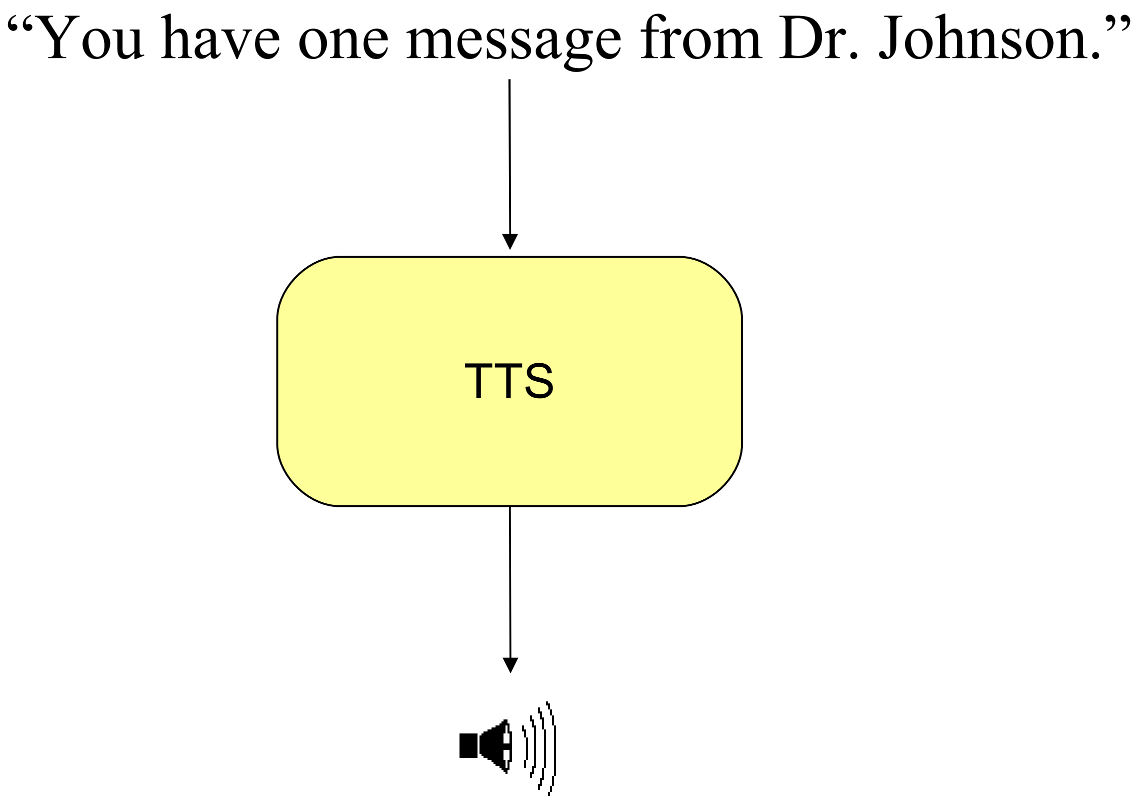
Foundations of Language Science and Technology
Speech synthesis

Marc Schröder, DFKI
schroed@dfki.de

21.12.2005

What is text-to-speech synthesis?



“You have one message from Dr. Johnson.”

TTS

Applications of TTS

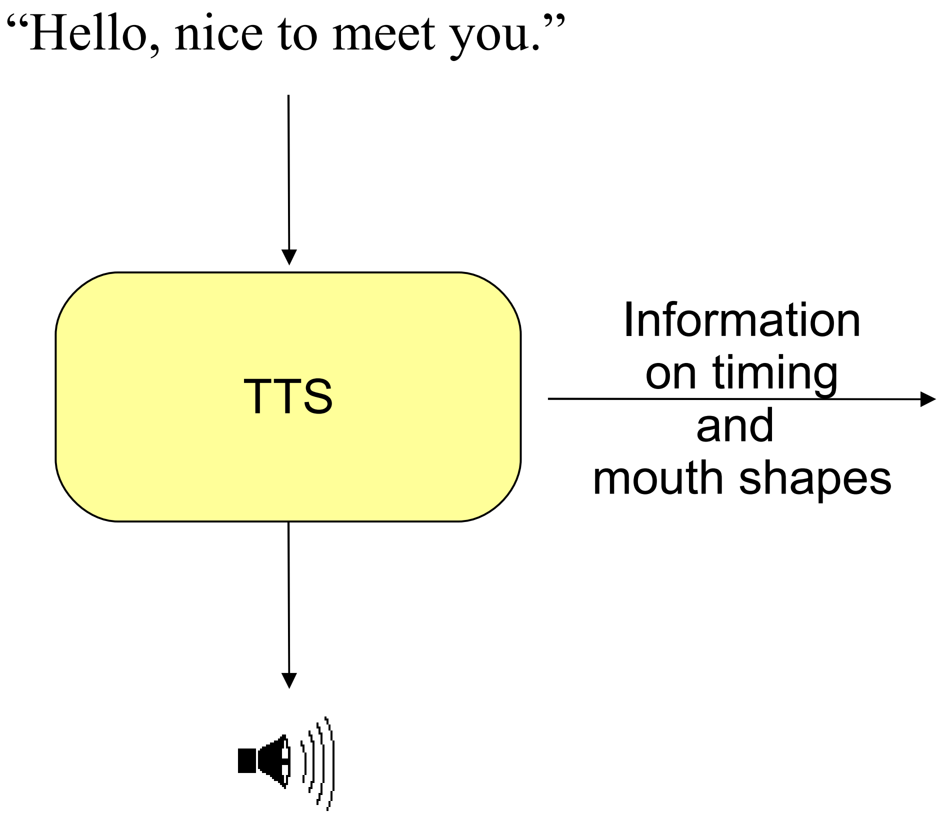
- Texts readers
  - for the blind
  - in eyes-free environments (e.g., while driving)
- Telephone-based voice portals
- Multi-modal interactive systems
  - talking heads
  - “embodied conversational agents” (ECAs)

Telephone-based voice portals

Example: Synthesising a phone number

🔈	**monotonous**	0-6-8-1-3-0-2-5-3-0-3
🔈	**unnatural (SMS-to-speech example)**	0. 6. 8. 1. 3. 0. 2. 5. 3. 0. 3.
🔈	**optimal (Baumann & Trouvain, 2001)**	0681 - 302 - 53 - 03

A Talking Head



“Hello, nice to meet you.”

TTS

Information on timing and mouth shapes

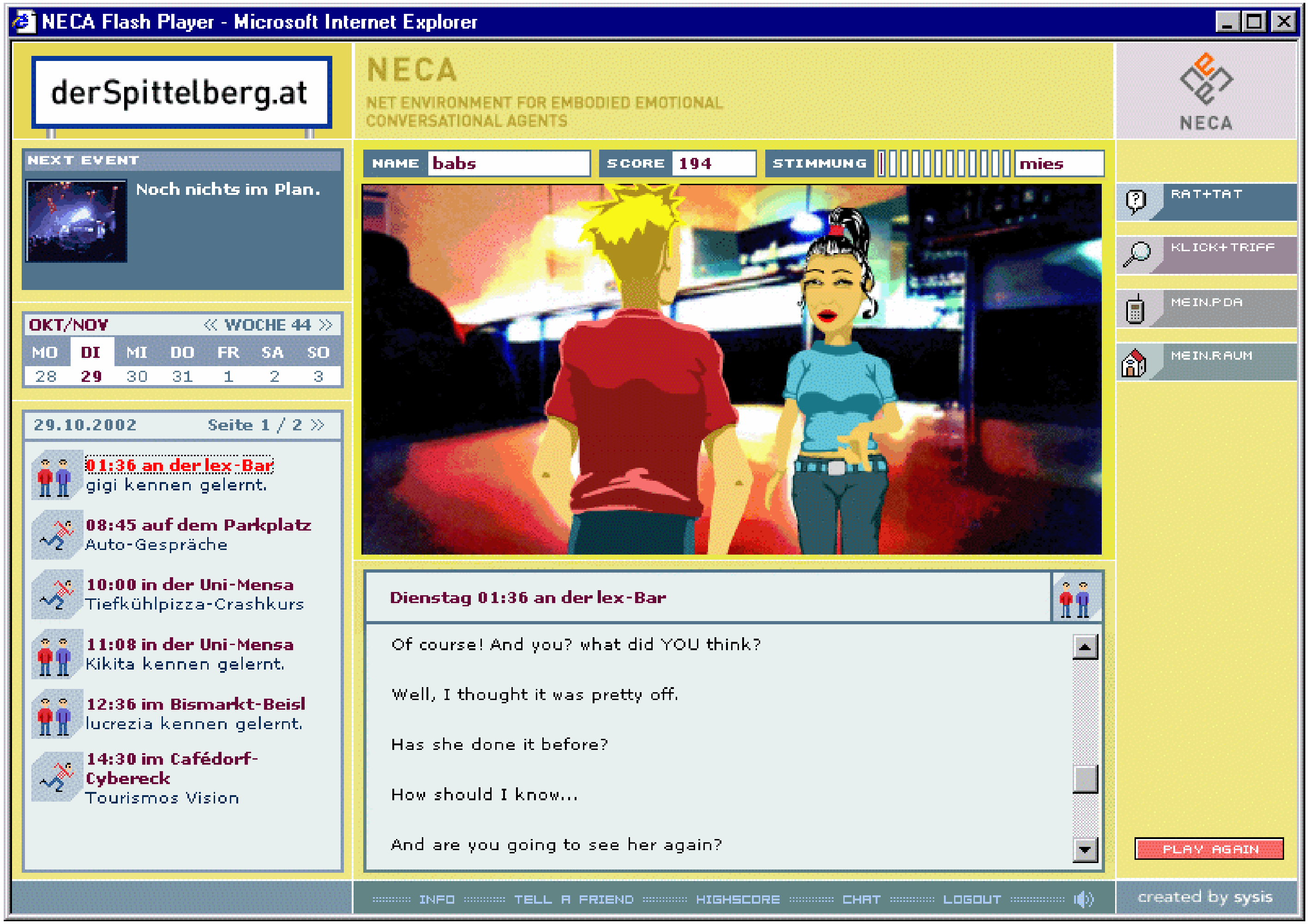
Facial Animation Model,
Computer Graphics Group,
MPI Saarbrücken

Web applications with ECAs

NECA eShowroom

www.eshowroom.org

Web applications with ECAs
NECA Socialite



NECA Flash Player - Microsoft Internet Explorer

derSpittelberg.at

NECA

NET ENVIRONMENT FOR EMBODIED EMOTIONAL CONVERSATIONAL AGENTS

NEXT EVENT

Noch nichts im Plan.

OKT/NOV « WOCHE 44 »

MO	DI	MI	DO	FR	SA	SO
28	29	30	31	1	2	3

29.10.2002 Seite 1 / 2 »

01:36 an der lex-Bar
gigi kennen gelernt.

08:45 auf dem Parkplatz
Auto-Gespräche

10:00 in der Uni-Mensa
Tiefkühlpizza-Crashkurs

11:08 in der Uni-Mensa
Kikita kennen gelernt.

12:36 im Bismarkt-Beisl
lucrezia kennen gelernt.

14:30 im Cafédorf-Cybereck
Tourismos Vision

NAME babs SCORE 194 STIMMUNG mies

Dienstag 01:36 an der lex-Bar

Of course! And you? what did YOU think?

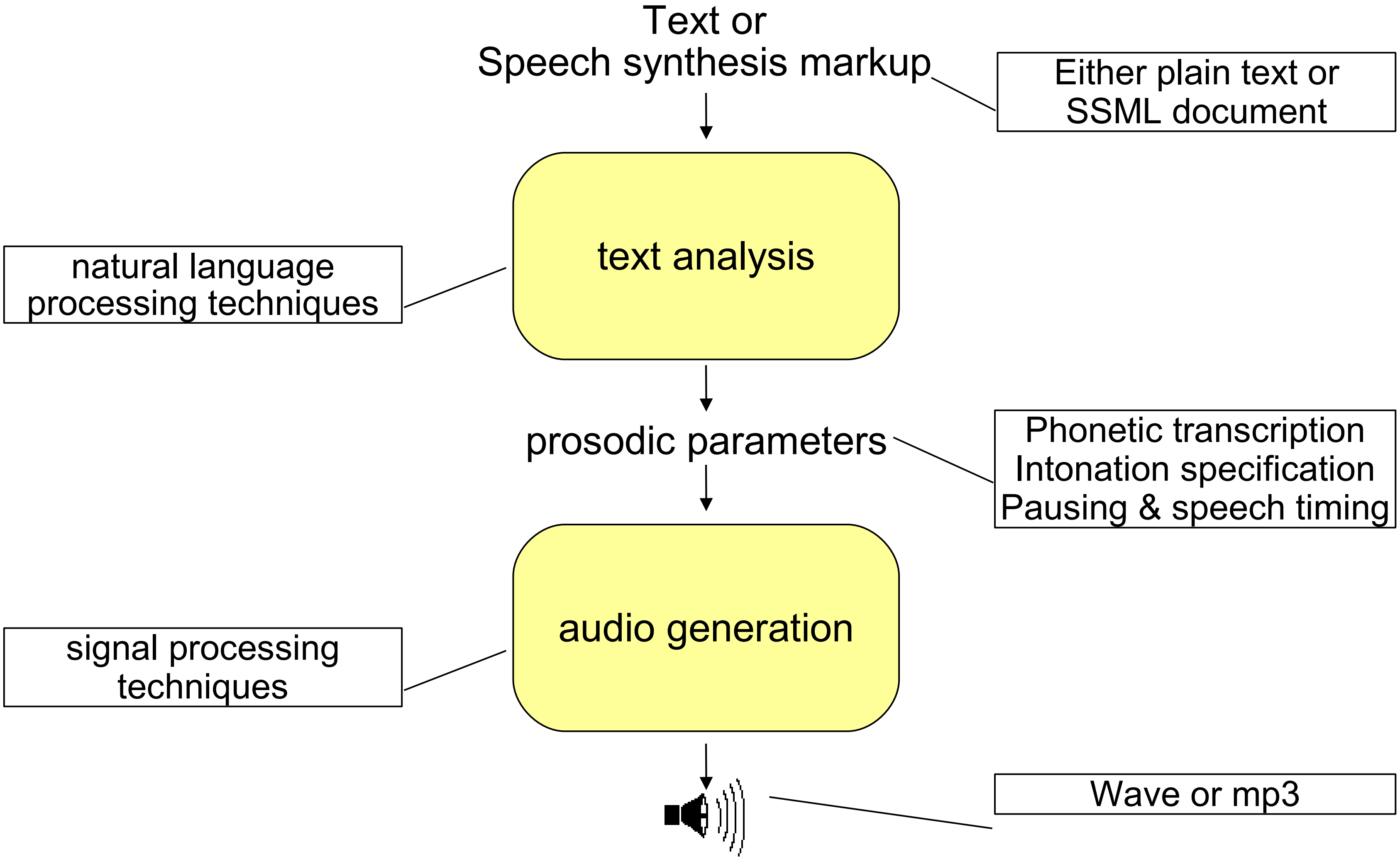
Well, I thought it was pretty off.

Has she done it before?

How should I know...

And are you going to see her again?

Structure of a TTS system



Text or
Speech synthesis markup

Either plain text or
SSML document

text analysis

natural language
processing techniques

prosodic parameters

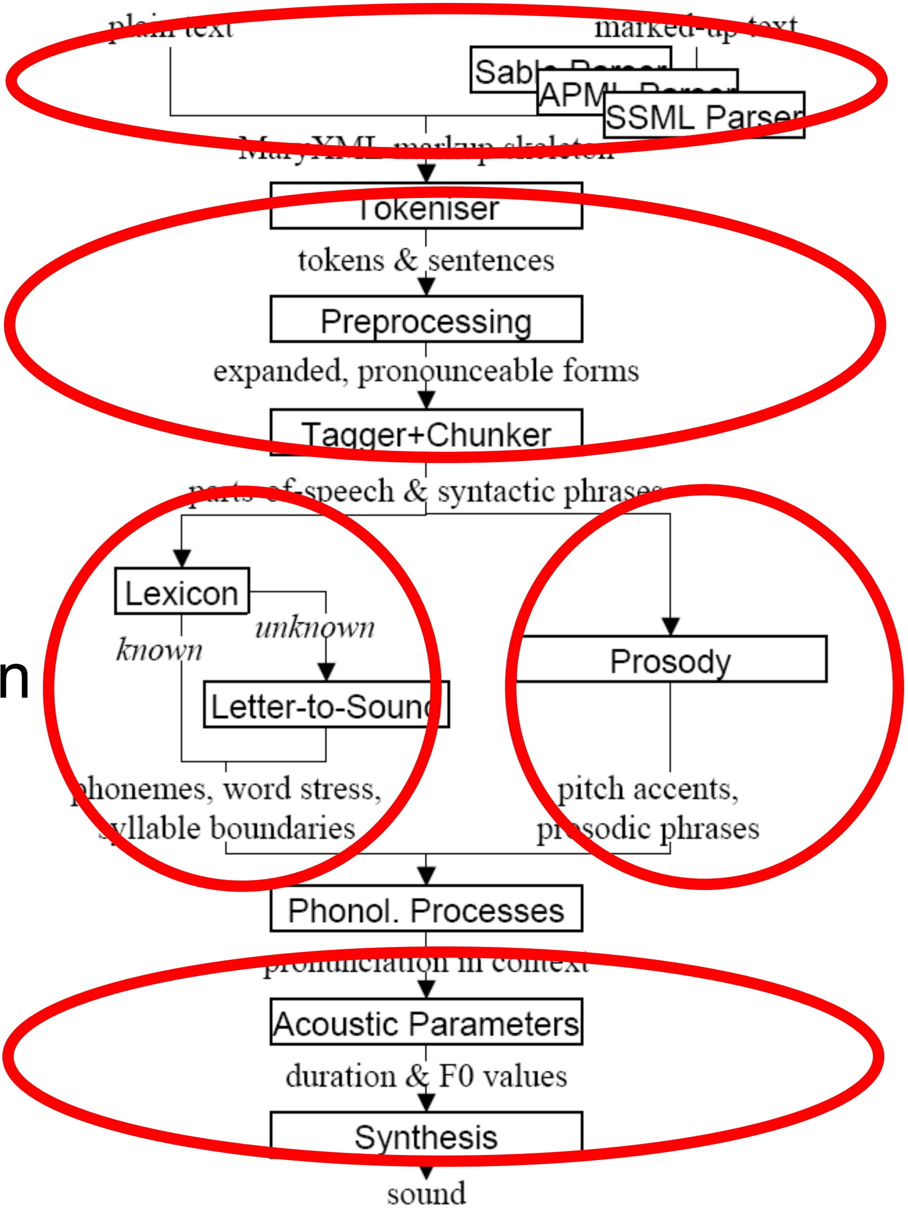
Phonetic transcription
Intonation specification
Pausing & speech timing

audio generation

signal processing
techniques

Wave or mp3

Structure of a TTS system: MARY



Input markup parser

plain text

marked-up text

Sable Parser

APML Parser

SSML Parser

MaryXML markup skeleton

Tokeniser

tokens & sentences

Shallow NLP

Preprocessing

expanded, pronounceable forms

Tagger+Chunker

parts-of-speech & syntactic phrases

Phonemisation

Lexicon

known

unknown

Letter-to-Sound

phonemes, word stress, syllable boundaries

Prosody

pitch accents, prosodic phrases

Phonol. Processes

pronunciation in context

Physical realisation

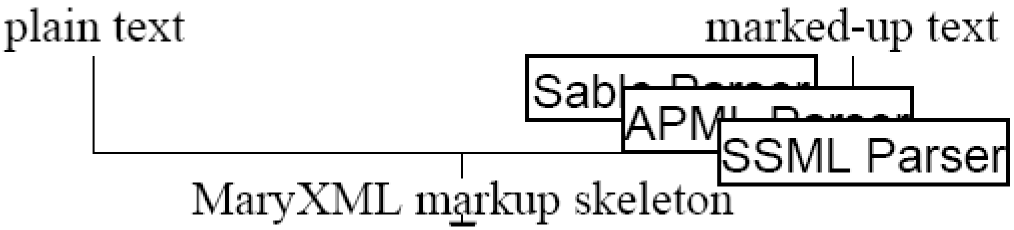
Acoustic Parameters

duration & F0 values

Synthesis

sound

System structure: Input markup parser



- System-internal XML representation **MaryXML**
- => speech synthesis markup parsing is simple XML transformation
- Use XSLT => easily adaptable to new markup language

Speech Synthesis Markup: SSML

- **Author (human or machine) provides additional information to the speech synthesis engine:**

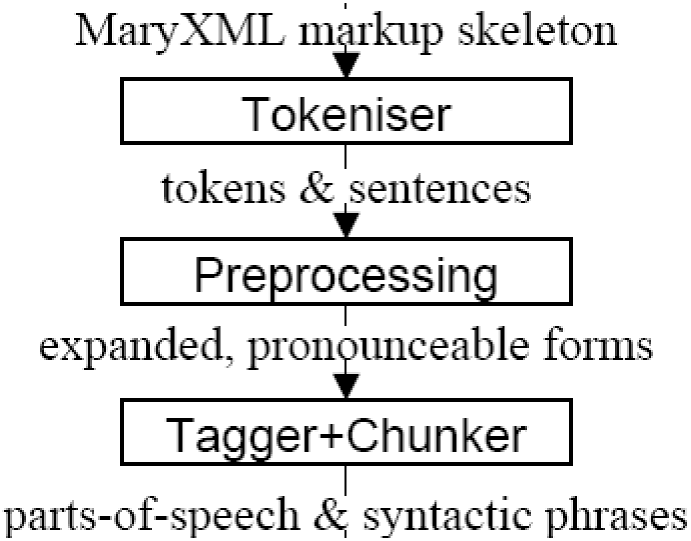
```
Er hat sich in München <emphasis> verlaufen </emphasis>
```

```
Im Jahr <say-as type=”date”> 1999 </say-as> wurden
<say-as type=”number:cardinal”> 1999 </say-as> Aufträge zur
Bestellnummer <say-as type=”number:digits”> 1999 </say-as>
erteilt.
```

```
<prosody pitch=”high” rate=”fast”>
Das müssen wir ganz schnell in Ordnung bringen!
</prosody>
```

```
<prosody pitch=”low” rate=”slow”>
Immer mit der Ruhe!
<prosody>
```

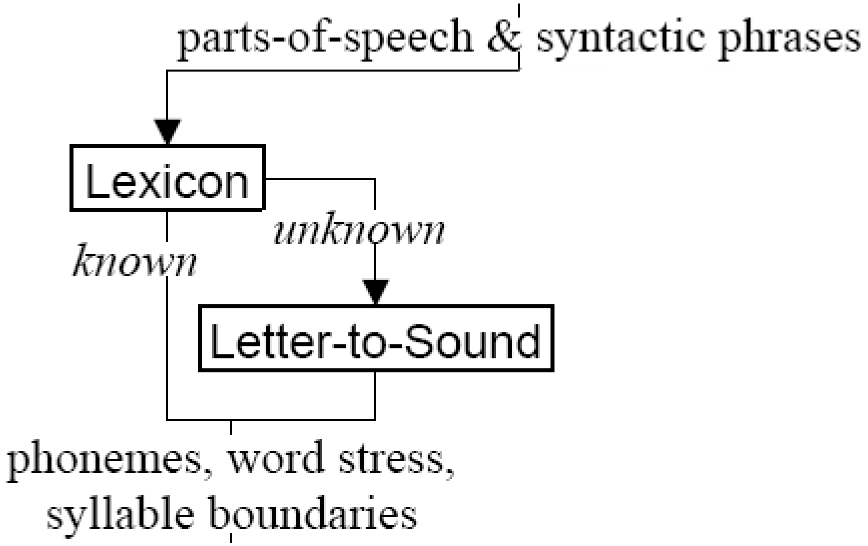
System structure: Shallow NLP



Preprocessing / Text normalisation

Net patterns (email, web addresses)	schroed@dfki.de
Date patterns	23.07.2001
Time patterns	12:24 h, 12:24 Uhr
Duration patterns	12:24 h, 12:24 Std.
Currency patterns	12,95 €
Measure patterns	123,09 km
Telephone number patterns	0681/302-5303
Number patterns (cardinal, ordinal, roman)	3 3. III
Abbreviations	engl.
Special characters	&

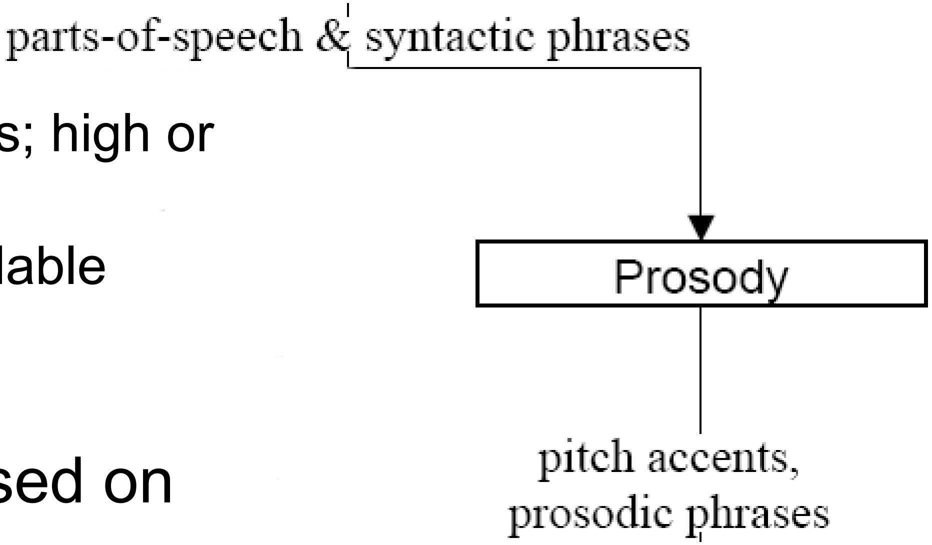
System structure: Phonemisation



- lexicon lookup
- letter-to-sound conversion
  - morphological decomposition
  - letter-to-sound rules
  - syllabification
  - word stress assignment

System structure: Prosody

- “Prosody”
  - intonation (accented syllables; high or low phrase boundaries)
  - rhythmic effects (pauses, syllable durations)
  - loudness, voice quality
- assign prosody by rule, based on
  - punctuation
  - part-of-speech
- modelled using “Tones and Break Indices” (ToBI)
  - tonal targets: accents, boundary tones
  - phrase breaks



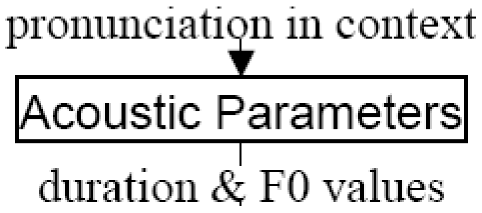
Prosody and meaning

Example: contrast and accentuation

- No, I said it's a blue MOON (not a blue horse)
- No, I said it's a BLUE moon (not a yellow moon)

- **Prosody can express contrast**
- **getting it wrong will make communication more difficult**

System structure: Calculation of acoustic parameters



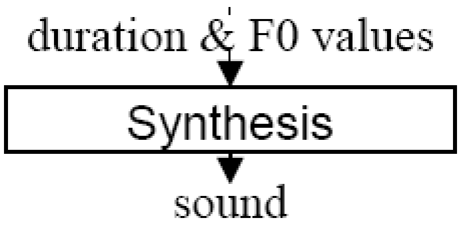
pronunciation in context

Acoustic Parameters

duration & F0 values

- timing:
  - segment duration predicted by rules or decision trees
- intonation:
  - rules from symbolic tone labels to F0-time targets

System structure: Waveform synthesis



Creating sound: Waveform synthesis technologies

- Formant synthesis
  - acoustic model of speech
  - generate acoustic structure by rule
  - robotic sound
- Concatenative synthesis
  - diphone synthesis
    - glue pre-recorded “diphones” together
    - adapt prosody through signal processing
  - unit selection synthesis
    - glue units from a large corpus of speech together
    - prosody comes from the corpus, (nearly) no signal processing

German TTS systems: Comparing signal quality

- **unit selection systems:**

L&H RealSpeak 🔊

AT&T Natural Voices 🔊 🔊

Loquendo ACTOR 🔊

Rhetorical rVoice 🔊

- **diphone systems:**

Elan TTS 🔊

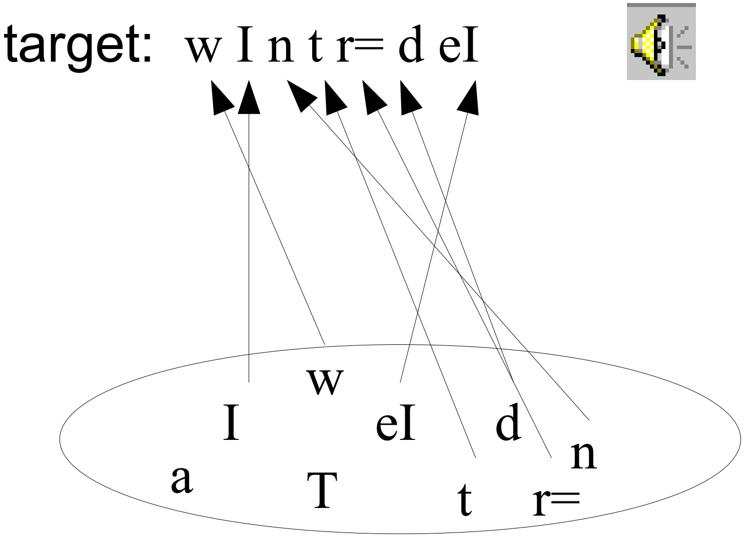
MBROLA-based (Infovox 🔊 , MARY 🔊 , Festival 🔊 , ...)

- **formant synthesis systems:**

SpeechWorks 🔊

Infovox 🔊

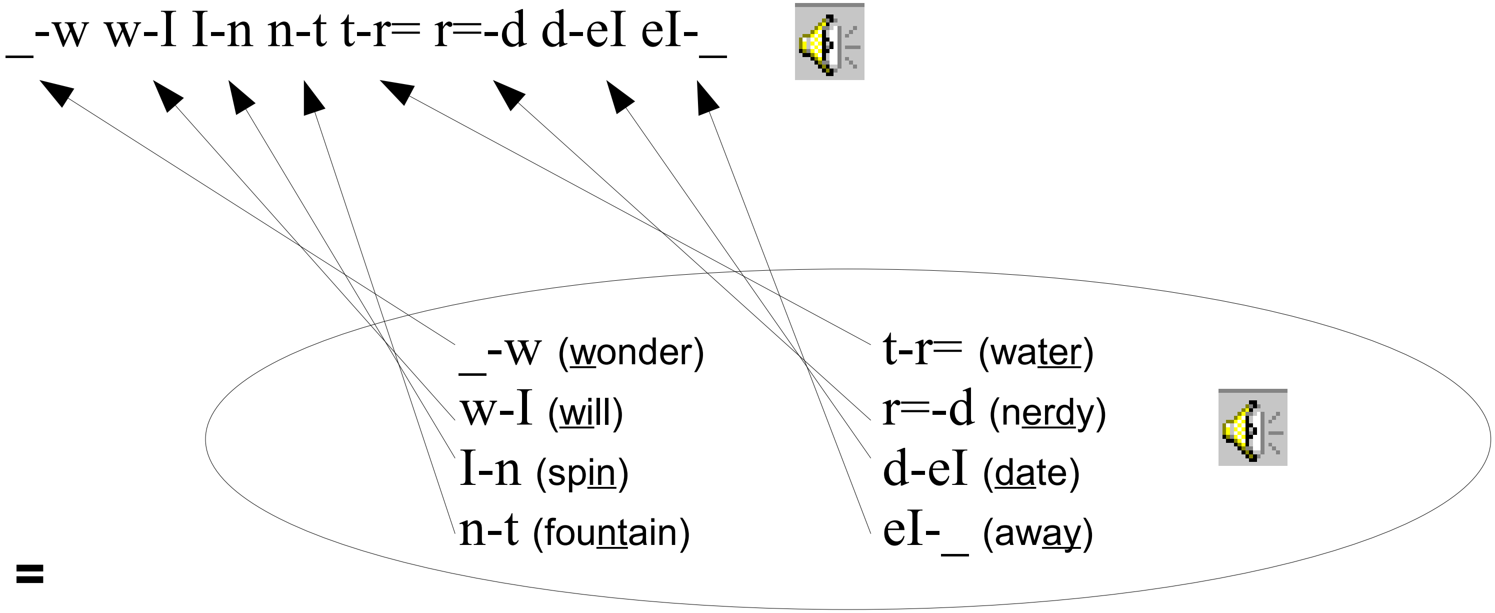
Concatenative synthesis: Isolated phones don't work



acoustic unit database
(units = **phone segments** recorded in isolation)

Concatenative synthesis: Diphones

target: w I n t r= d eI



_-w w-I I-n n-t t-r= r=-d d-eI eI-_

_-w (wonder)
w-I (will)
I-n (spin)
n-t (fountain)

t-r= (water)
r=-d (nerdy)
d-eI (date)
eI-_ (away)

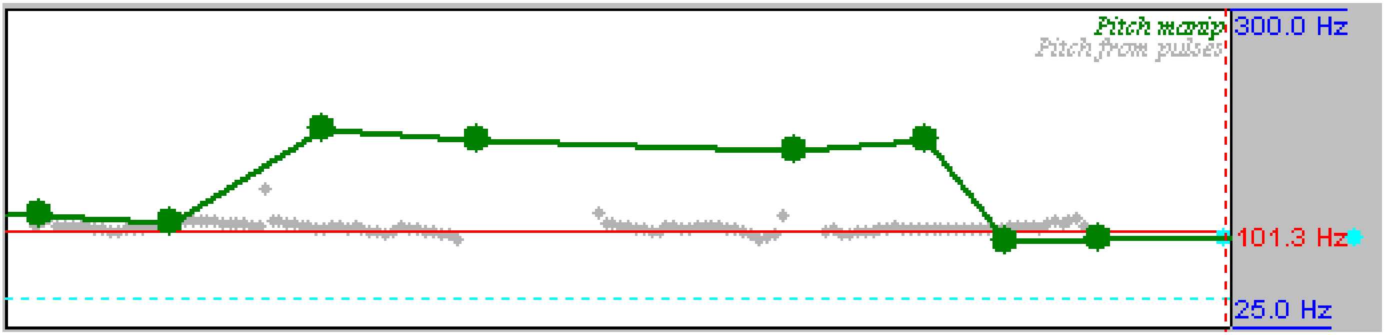
Diphones =
sound segments
from the middle of one phone
to the middle of the next phone

acoustic unit database
units = **diphone segments**
recorded in carrier words
(flat intonation)

Concatenative synthesis: Diphones (2)

target: w I n t r= d eI

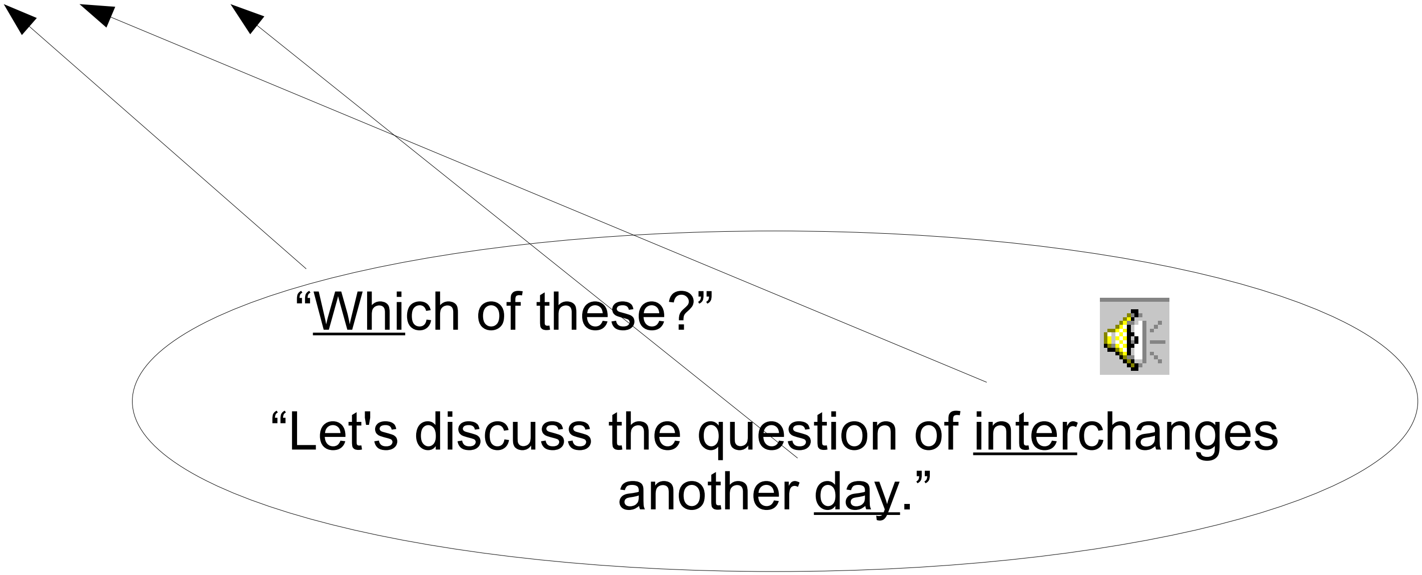
_-w w-I I-n n-t t-r= r=-d d-eI eI-_



PSOLA
pitch
manipulation

Concatenative synthesis
Unit selection

target: w I n t r= d eI



“Which of these?”

“Let's discuss the question of interchanges another day.”

acoustic unit database
units = **(di-)phone segments** recorded in natural sentences (natural intonation)

Emotional / Expressive TTS

Expressive speech synthesis
Formant synthesis

- Acoustic modelling of speech
- Many degrees of freedom, can potentially reproduce speech perfectly
- Rule-based formant synthesis: Imperfect rules for acoustic realisation of articulation
  => robot-like sound

Examples:

Janet Cahn (1990):
angry
happy
sad
fearful

Felix Burkhardt (2001):
neutral
angry
happy
sad
fearful

Diphone synthesis

- Diphones = small units of recorded speech
  - from middle of one sound to middle of next sound
  - e.g. [grEIt] = _-g g-r r-EI EI-t t-_
- Signal manipulation to force pitch (F0) and duration into a target contour
  - Can control prosody, but not voice quality

Examples:	neutral		
	angry		angry
Marc Schröder (1999):	happy	Ignasi Iriondo (2004):	happy
	sad		sad
	fearful		fearful

Diphone synthesis

- Is voice quality indispensable?
  - Interesting diversity of opinions in the literature
  - Tentative conclusion: “It depends!”
    - ...on the emotion (Montero et al., 1999)
      - prosody conveys surprise, sadness
      - voice quality conveys anger, joy
    - ...on speaker strategies (Schröder, 1999)

angry1 orig_angry1 angry2 orig_angry2

Diphone synthesis

- Partial remedy: Record voice qualities
- Schröder & Grice (2003): Diphone databases with three levels of vocal effort

male:	loud	modal	soft
female:	loud	modal	soft

- Voice quality interpolation: Turk et al. (2005)

female: loud 1 2 modal 3 4 soft

- Not yet successful: smiling voice

modal1	smile1
modal2	smile2

Expressive speech synthesis
Unit selection synthesis

- Select small speech units out of very large speech corpus (e.g., 5 hours of speech)
- Avoid signal manipulation to maintain natural prosody from the units
  - Can**not** control prosody or voice quality
  - Very good “playback” quality with emotional recordings

Examples:

Akemi Iida (2000):	angry happy sad	Ellen Eide (IBM, 2004):	good news bad news

Technologies for expressive TTS: Summary

- “Explicit modelling” approaches
  - low naturalness
  - high flexibility, high control over acoustic parameters
  - explicit models of emotional prosody
- “Playback” approaches
  - high naturalness
  - no flexibility, no control over acoustic parameters
  - emotional prosody implicit in recordings