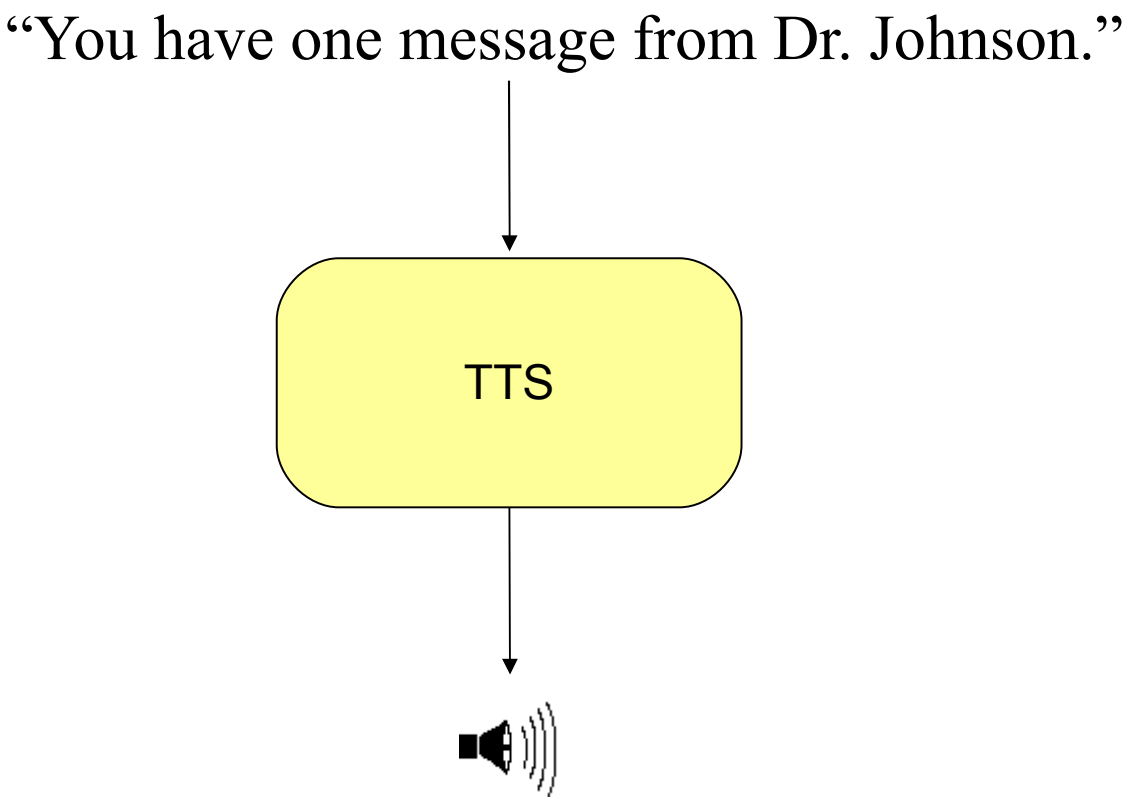
Foundations of Language Science and Technology

Speech synthesis

Marc Schröder, DFKI
schroed@dfki.de

28 January 2009

What is text-to-speech synthesis?



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

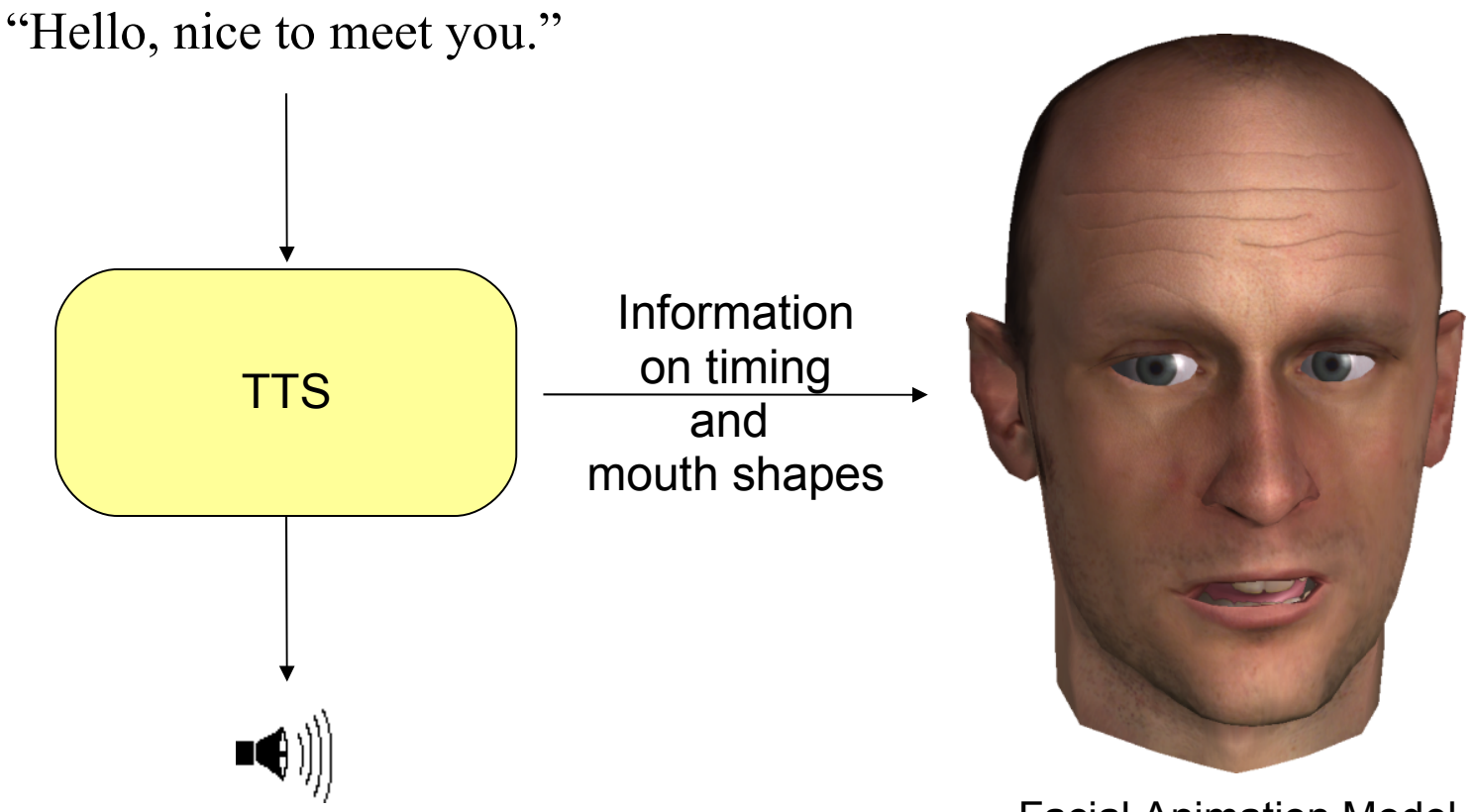
Example: Synthesising a phone number

- **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

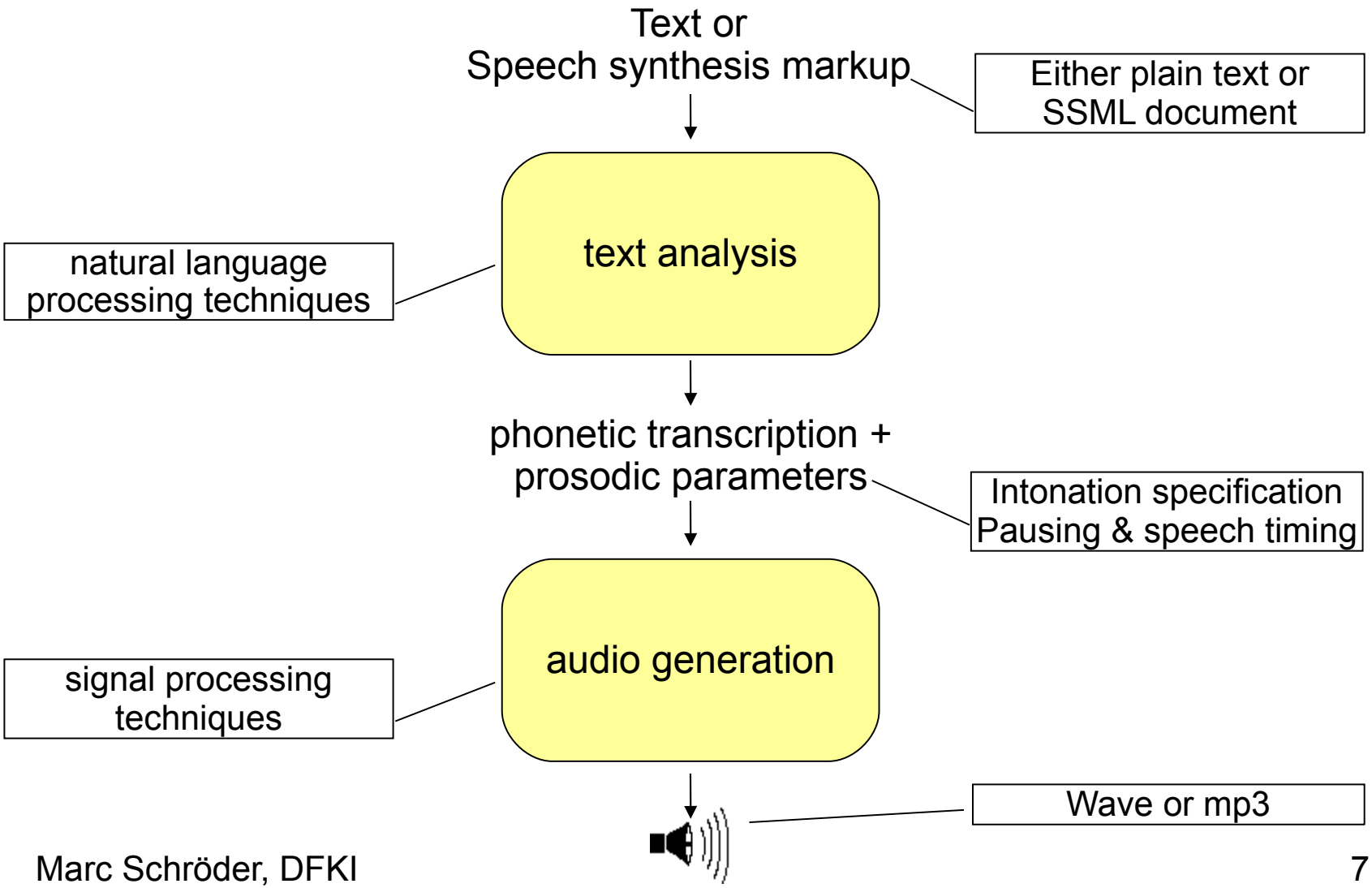


Facial Animation Model,
Computer Graphics Group,
MPI Saarbrücken

An instrumented Poker game: “AI Poker”

- user is playing against two virtual characters
  - user shuffles and deals (RFID)
- game events trigger emotions in characters
- emotion is expressed in synthetic voices

Structure of a TTS system



Structure of a TTS system: MARY

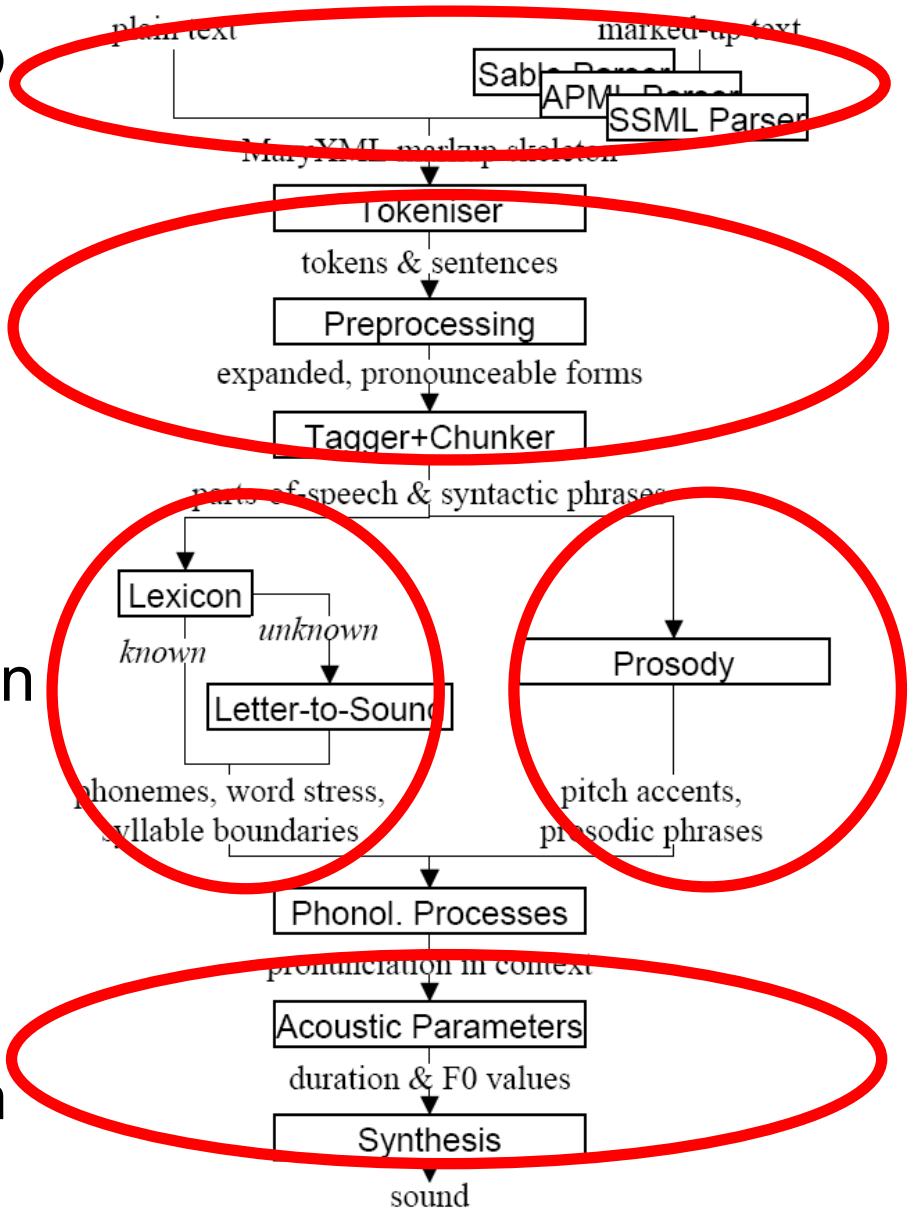
Input markup parser

Shallow NLP

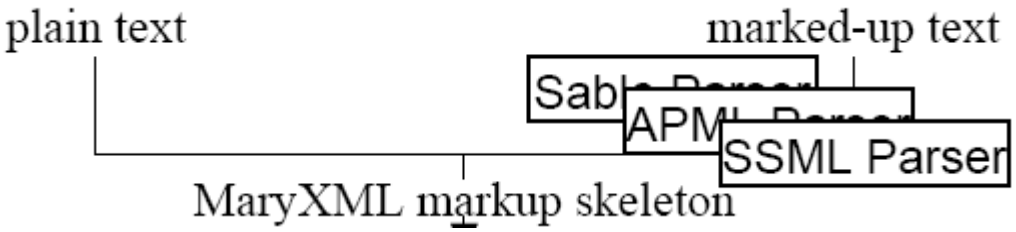
Phonemisation

Physical realisation

Prosody



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:**

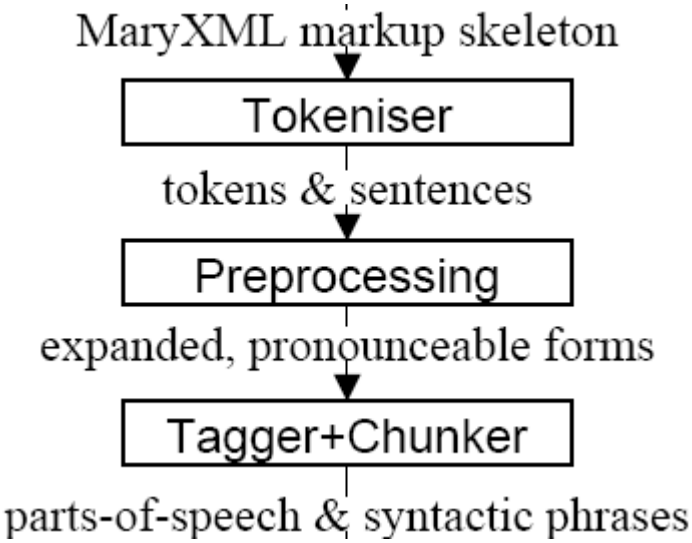
```
Er hat sich in München <emphasis> verlaufen </emphasis>
```

```
Im Jahr <say-as interpret-as="date" format="y">1999</say-as>
wurden <say-as interpret-as="cardinal">1999</say-as> Aufträge
zur Bestellnummer <say-as interpret-as="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>
```

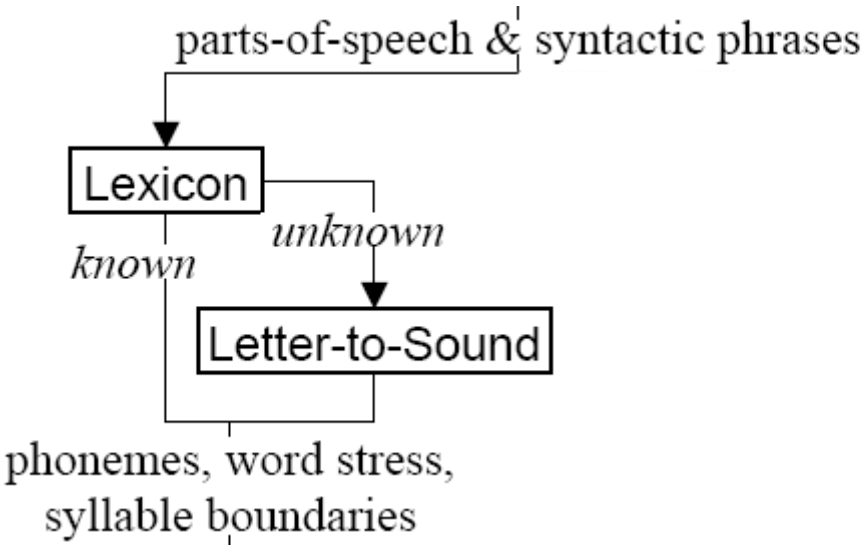
System structure: Shallow NLP



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	&

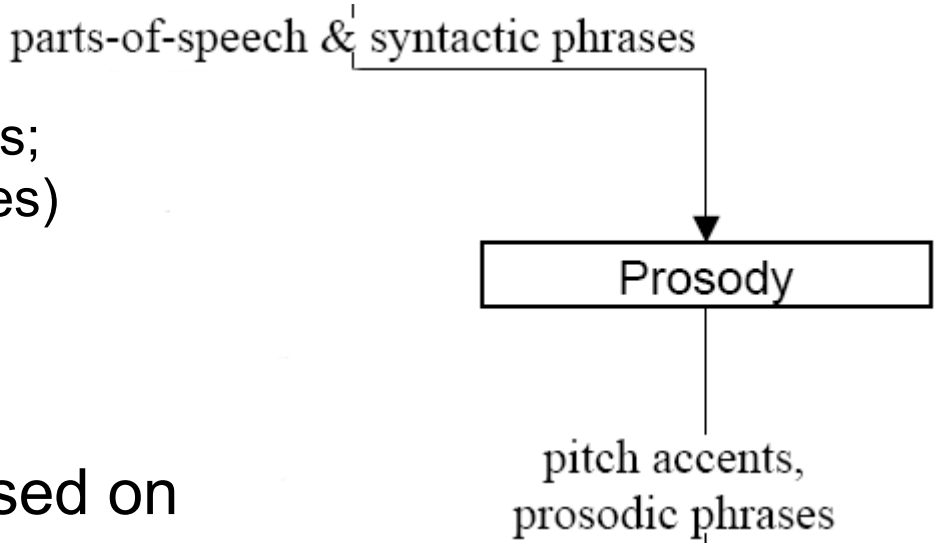
System structure: Phonemisation



- 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



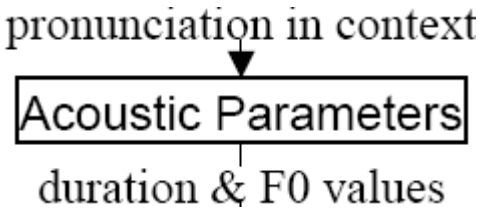
Prosody and meaning
Example: contrast and accentuation

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

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

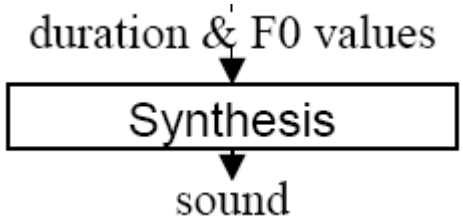
- **Prosody can express contrast**
- **getting it wrong will make communication more difficult**

System structure: Calculation of acoustic parameters



- timing:
  - segment duration predicted
    - by rules
    - or by decision trees
- intonation:
  - fundamental frequency curve predicted
    - by rules
    - or by decision trees

System structure: Waveform synthesis



Creating sound: Waveform synthesis technologies (1)

- Formant synthesis
  - acoustic model of speech
  - generate acoustic structure by rule
  - robotic sound

Creating sound: Waveform synthesis technologies (2)

- 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

Creating sound:
Waveform synthesis technologies (3)

- Statistical-parametric speech synthesis
  - with Hidden Markov Models
  - models trained on speech corpora
  - no data needed at runtime => small footprint

Examples of various speech synthesis systems

- **unit selection systems:**

L&H RealSpeak 

AT&T Natural Voices 

Loquendo ACTOR 

MARY 

- **diphone systems:**

Elan TTS 

MBROLA-based (MARY )

- **formant synthesis systems:**

SpeechWorks 

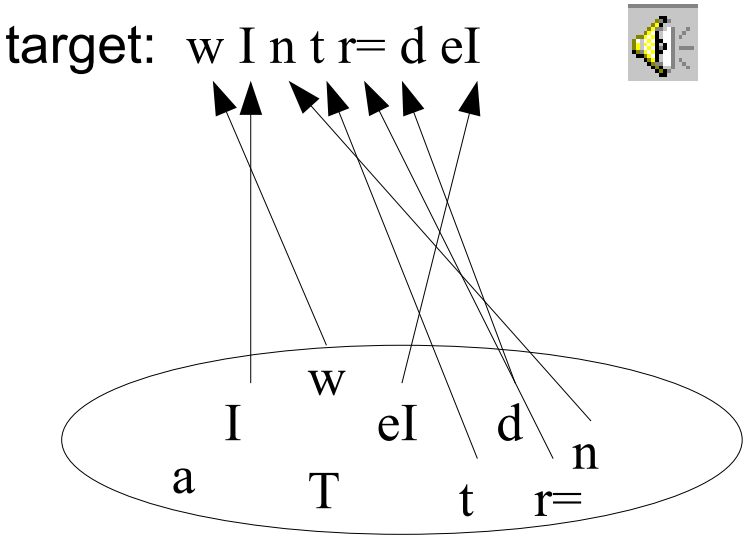
Infovox 

- **HMM-based systems:**

MARY  

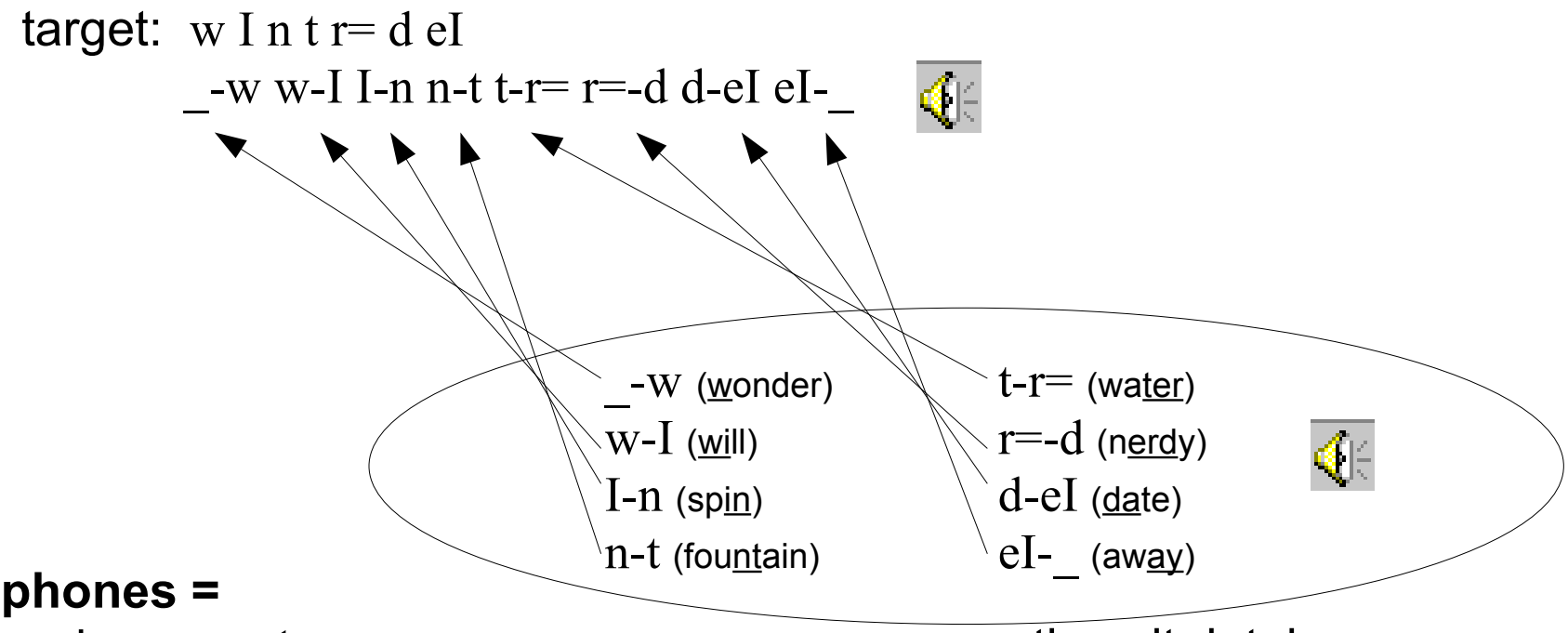
(others exist: HTS, USTC, Festival, ...)

Concatenative synthesis: Isolated phones don't work



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

Concatenative synthesis: Diphones



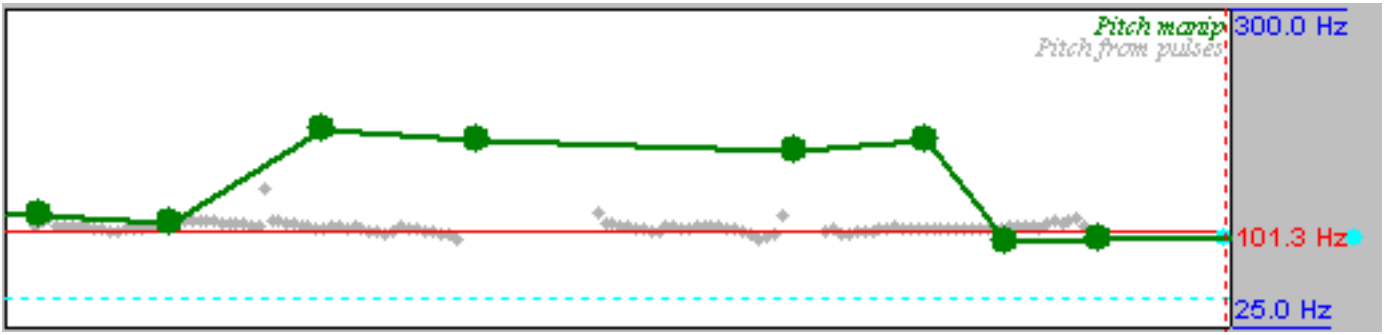
Diphones =
sound segments
from the middle of one phone
to the middle of the next phone

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

Concatenative synthesis: Diphones (2)

target: w I n t r= d eI

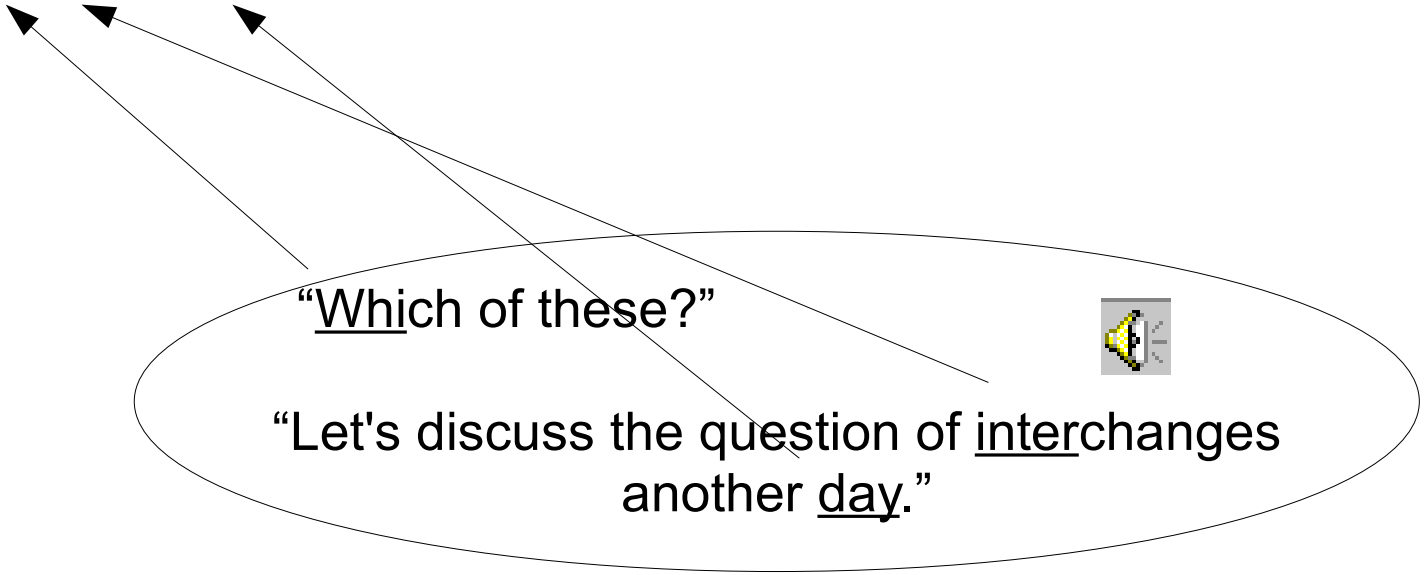
_-w w-I I-n n-t t-r= r=-d d-eI eI-_



PSOLA pitch manipulation

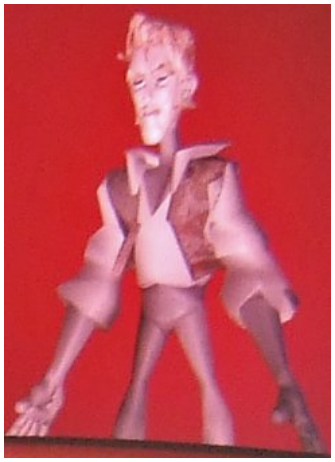
Concatenative synthesis
Unit selection

target: w I n t r= d eI



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

AI Poker: The voices of Sam and Max



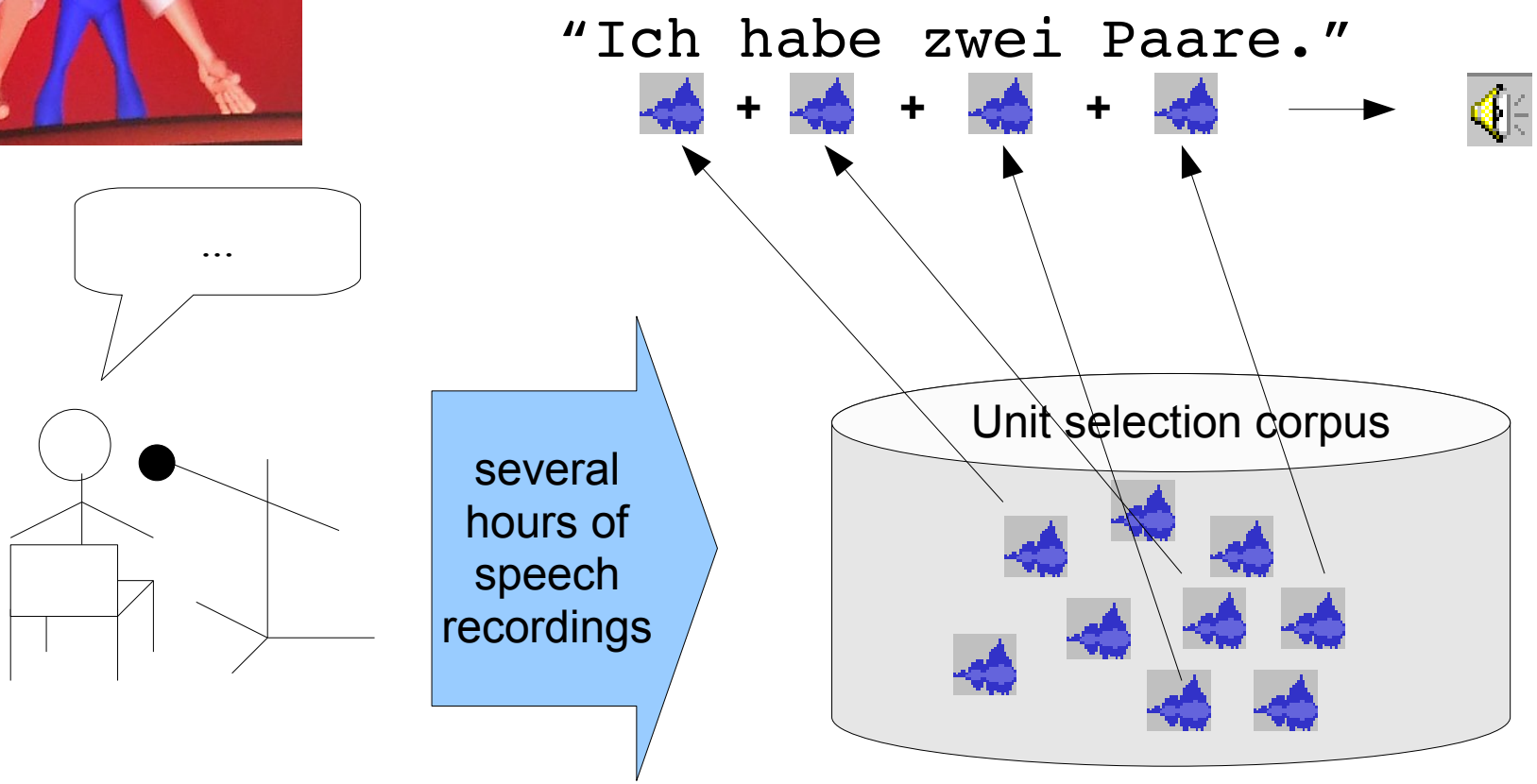
Sam:

- Unit Selection Synthesis
- Voice specifically recorded for AI Poker
- Natural sound within poker domain

Max:

- HMM-based synthesis
- Sound quality is limited but constant with any text

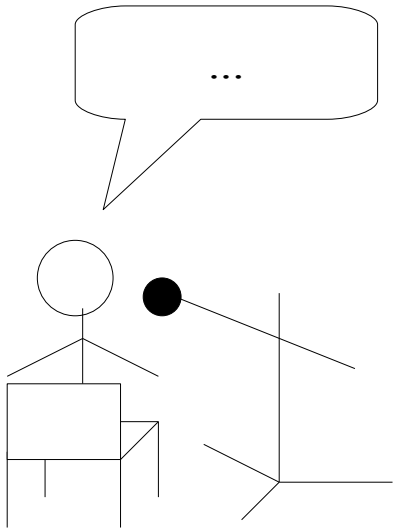
Sam's voice: Unit selection syntheis



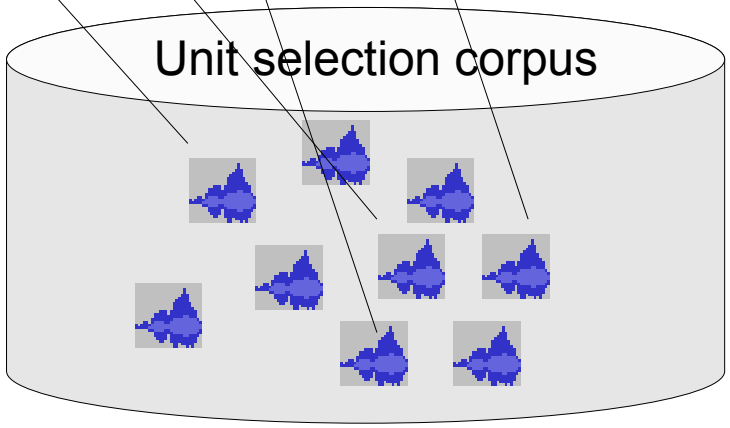
=> very good quality within the poker domain!

Sam's voice: Unit selection syntheis

"Ich kann auch ganz andere Sachen..."

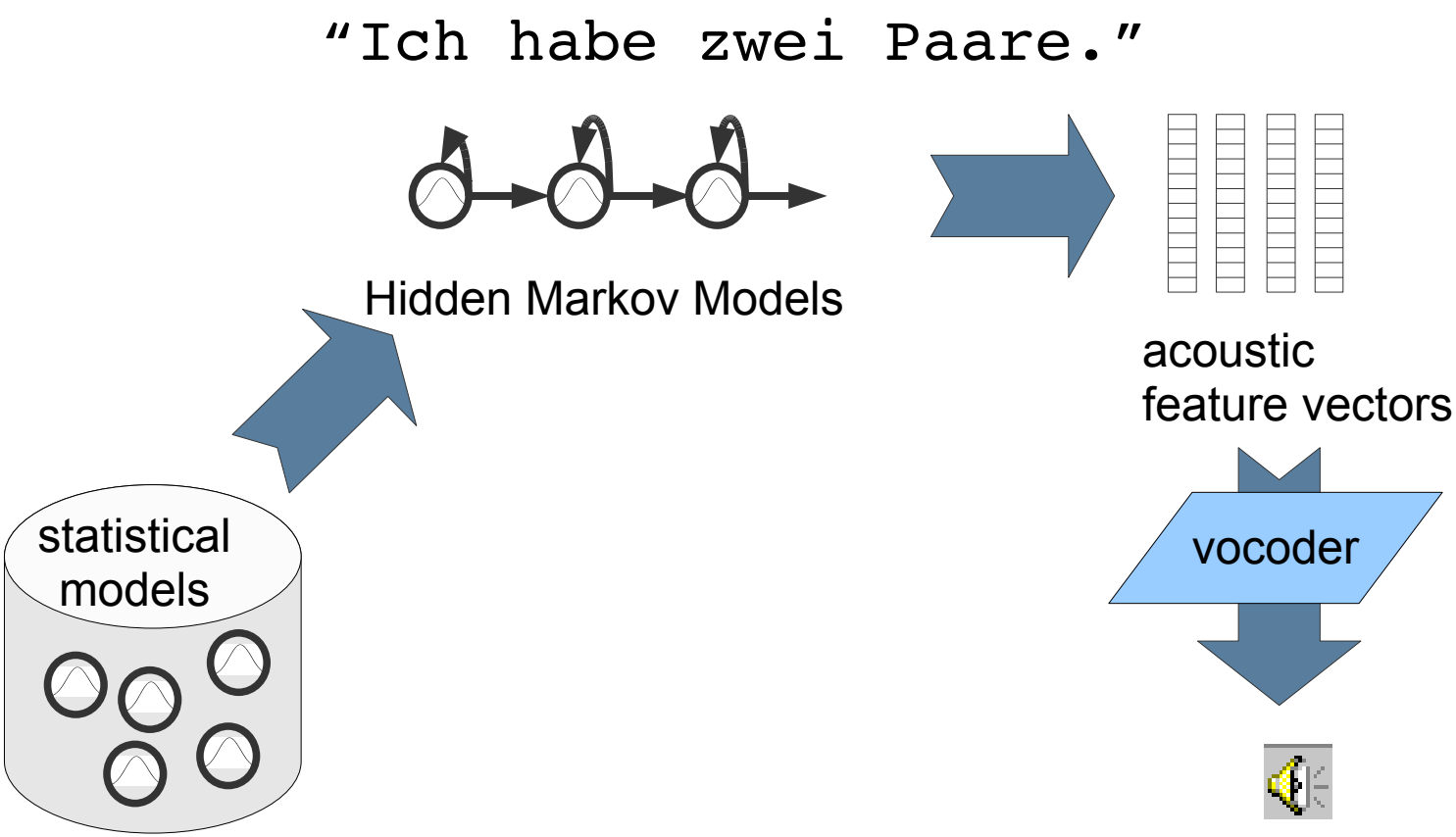


several hours of speech recordings



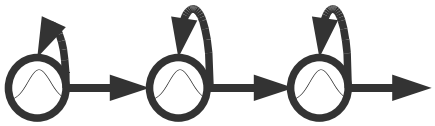
reduced quality with arbitrary text

Max's voice: HMM-based synthesis

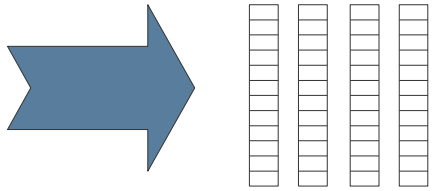


Max's voice: HMM-based synthesis

"Ich kann auch ganz andere Sachen..."

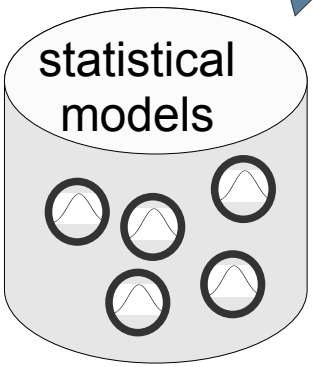


Hidden Markov Models



acoustic
feature vectors

vocoder



constant quality with arbitrary text