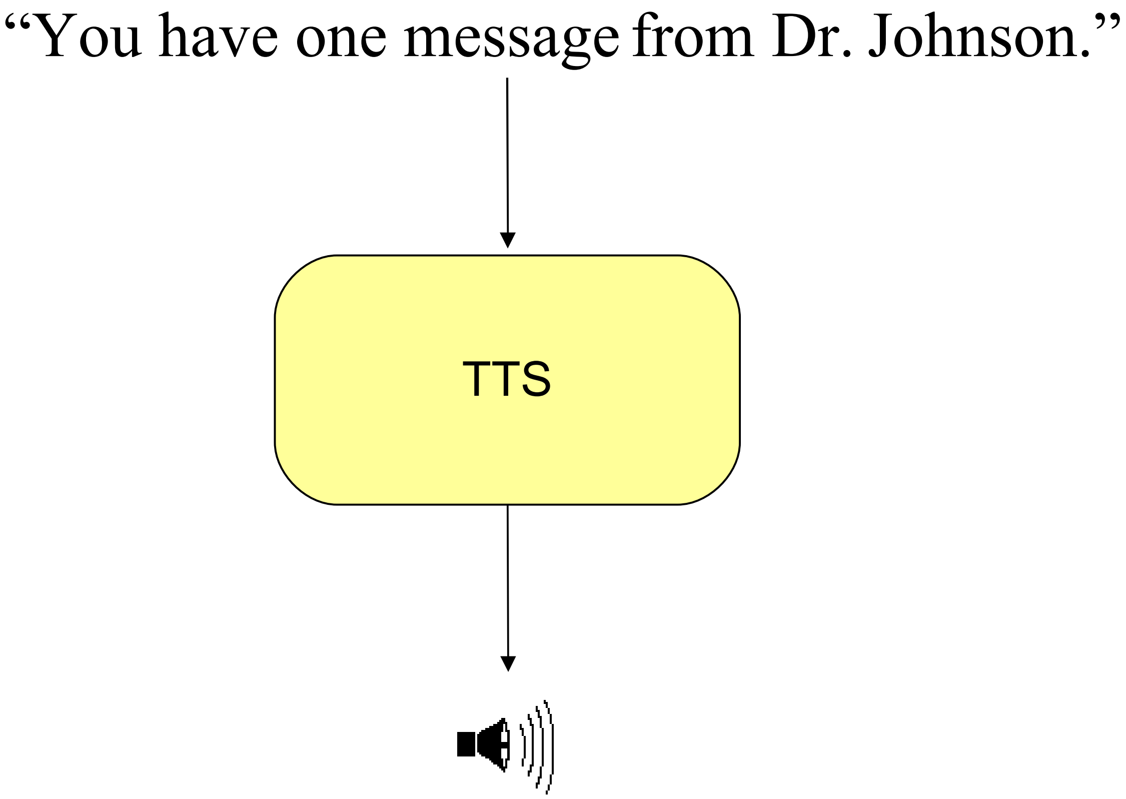
Foundations of Language Science and Technology
Speech synthesis

Marc Schröder, DFKI
schroed@dfki.de

06 February 2008

What is text-to-speech synthesis?



“You have one message from Dr. Johnson.”

TTS

Applications of TTS

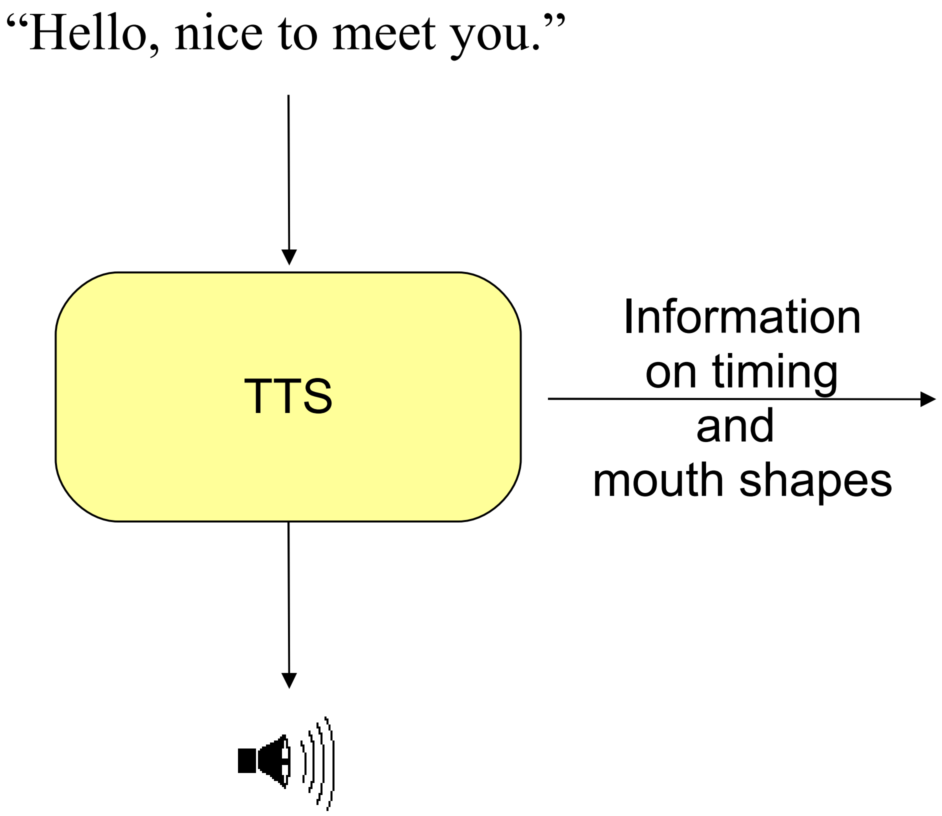
- Texts readers
  - for the blind
  - in eyes-free environments (e.g., while driving)
- Telephone-based voice portals
- Multi-modal interactive systems
  - talking heads
  - “embodied conversational agents” (ECAs)

Telephone-based voice portals

Example: Synthesising a phone number

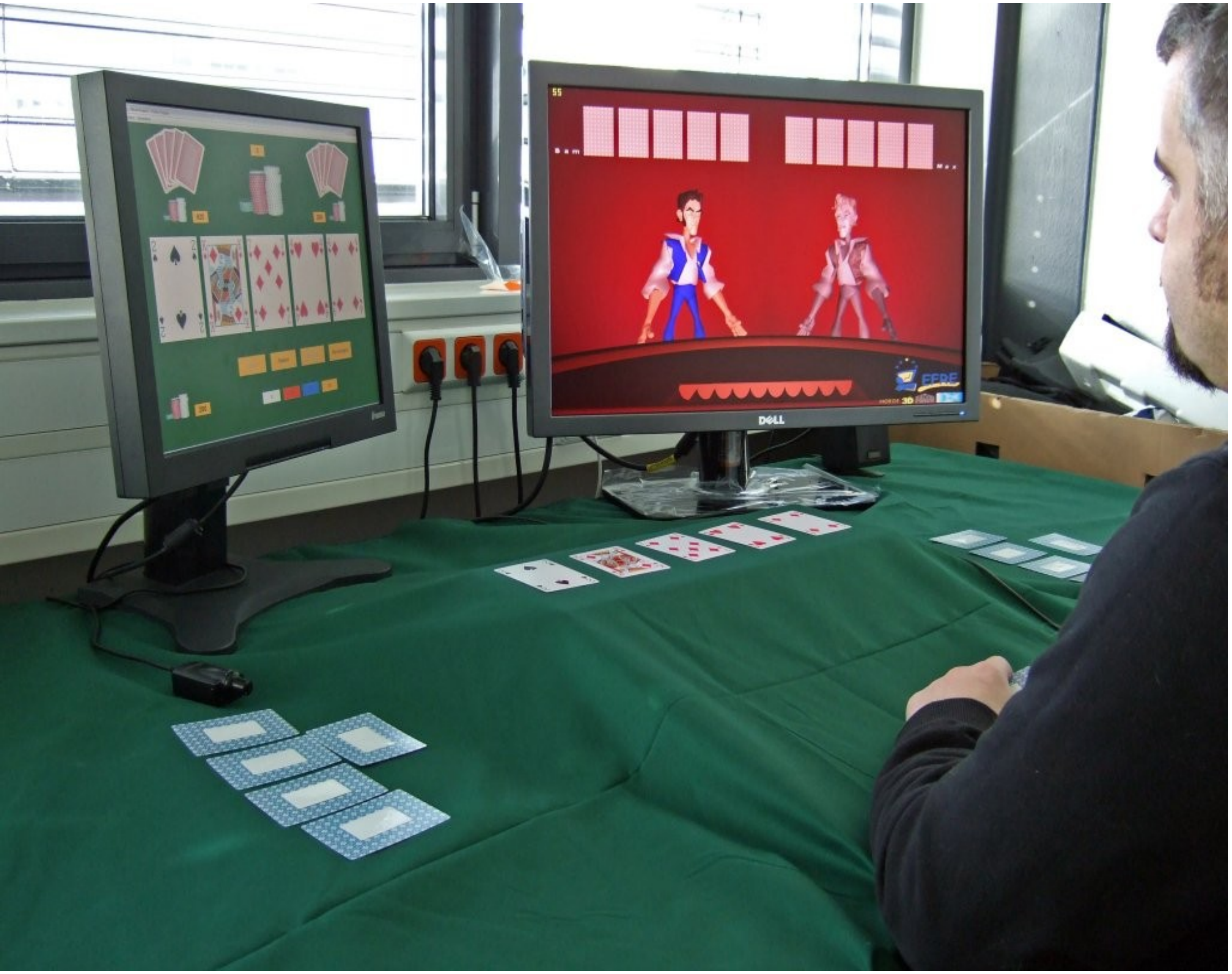
🔈	**monotonous**	0-6-8-1-3-0-2-5-3-0-3
🔈	**unnatural (SMS-to-speech example)**	0. 6. 8. 1. 3. 0. 2. 5. 3. 0. 3.
🔈	**optimal (Baumann & Trouvain, 2001)**	0681 - 302 - 53 - 03

A Talking Head


"Hello, nice to meet you."
TTS
Information on timing and mouth shapes

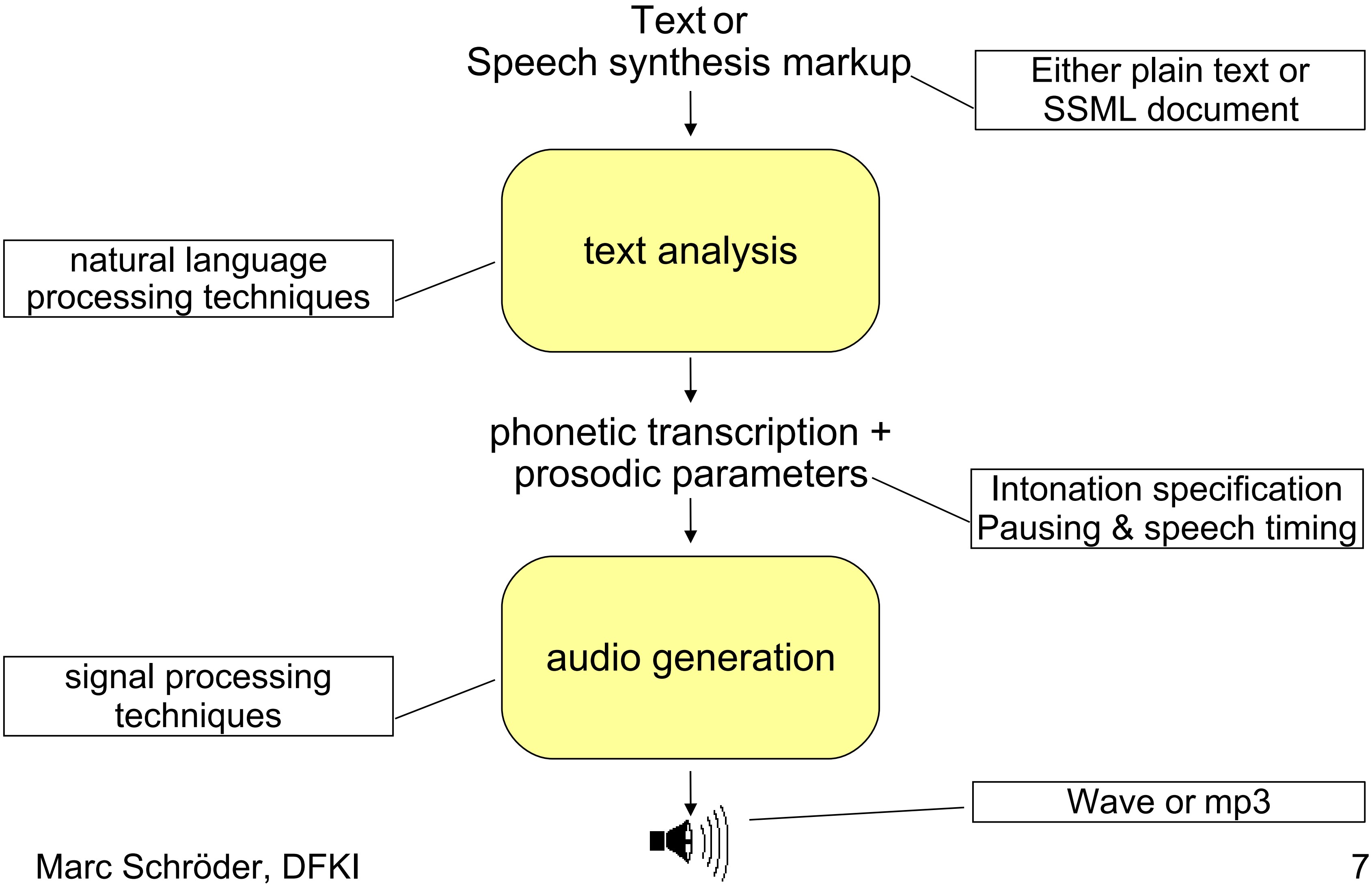
Facial Animation Model,
Computer Graphics Group,
MPI Saarbrücken

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



Text or
Speech synthesis markup

Either plain text or
SSML document

text analysis

natural language
processing techniques

phonetic transcription +
prosodic parameters

Intonation specification
Pausing & speech timing

audio generation

signal processing
techniques

Wave or mp3

Structure of a TTS system: MARY

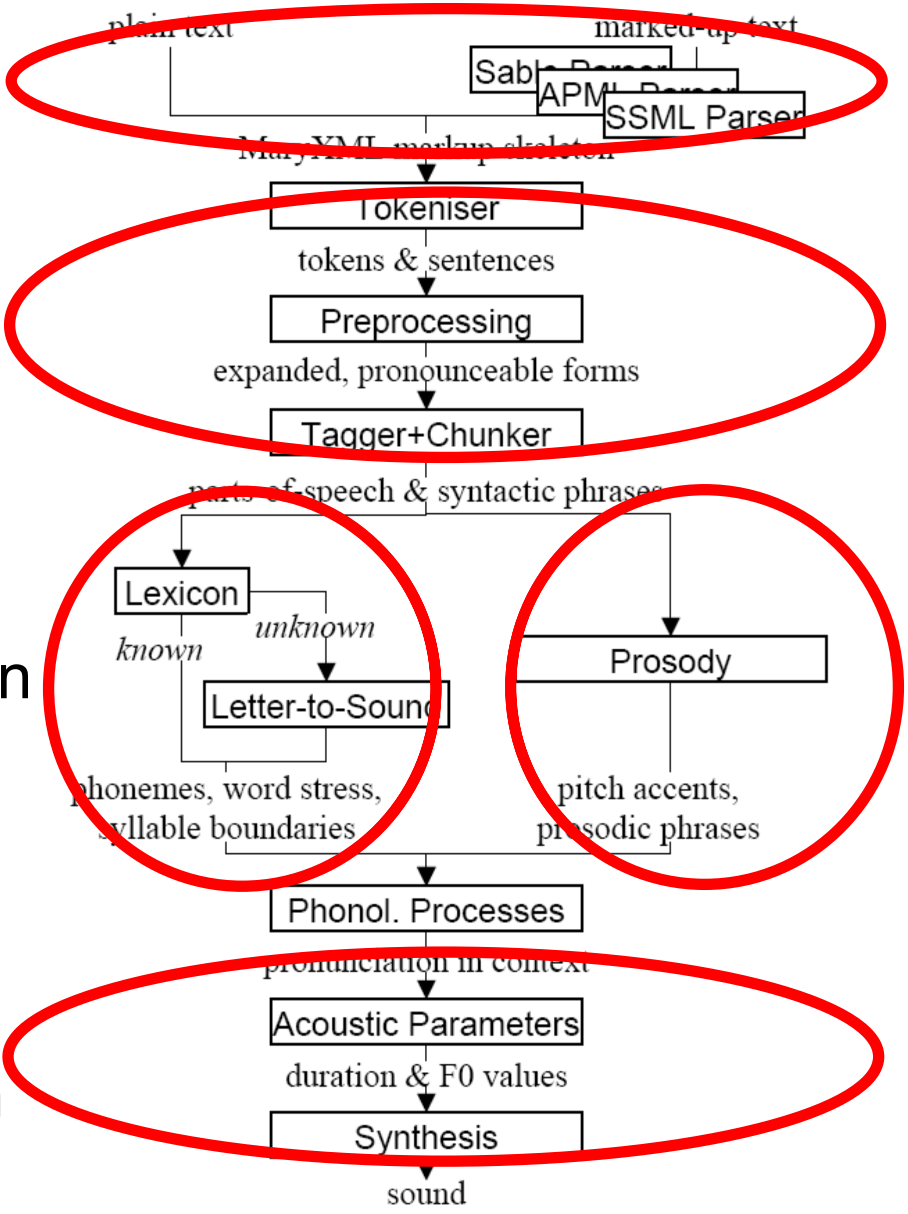
Input markup parser

Shallow NLP

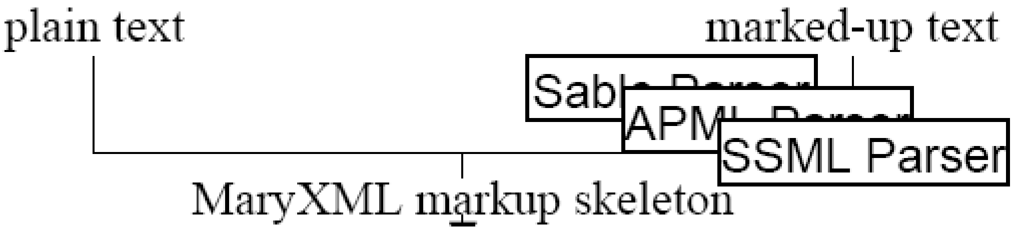
Phonemisation

Prosody

Physical realisation



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:**

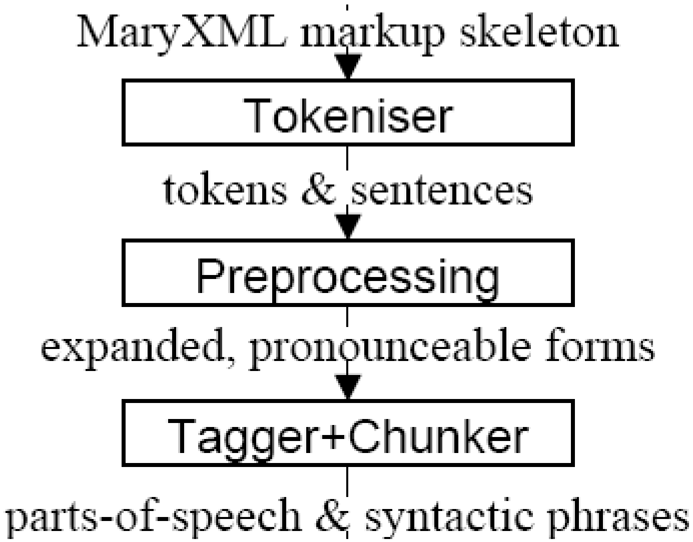
```
Er hat sich in München <emphasis> verlaufen </emphasis>
```

```
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



MaryXML markup skeleton

Tokeniser

tokens & sentences

Preprocessing

expanded, pronounceable forms

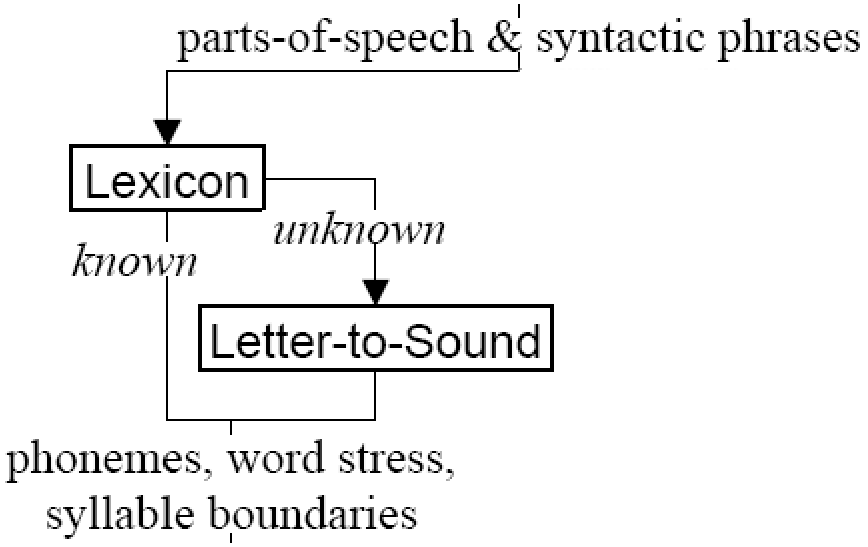
Tagger+Chunker

parts-of-speech & syntactic phrases

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	&

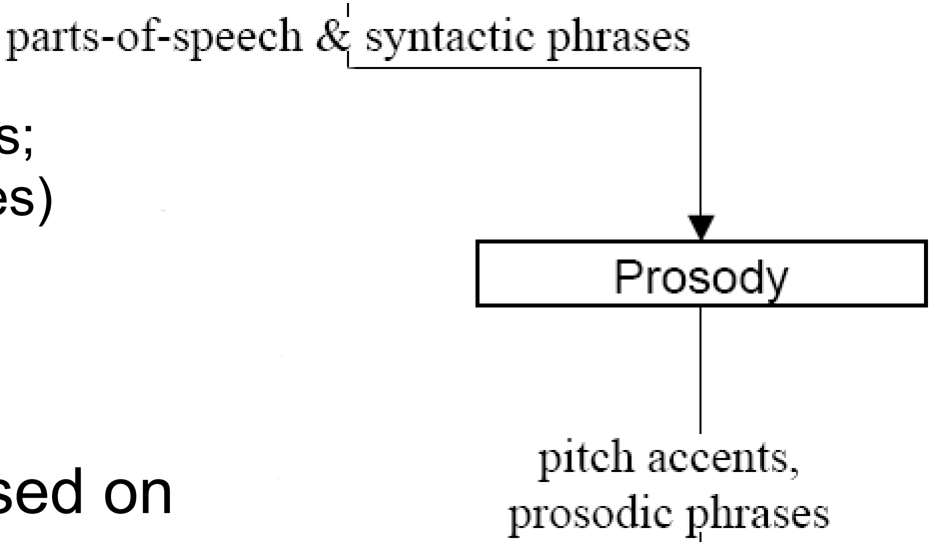
System structure: Phonemisation



- 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


parts-of-speech & syntactic phrases
Prosody
pitch accents, prosodic phrases

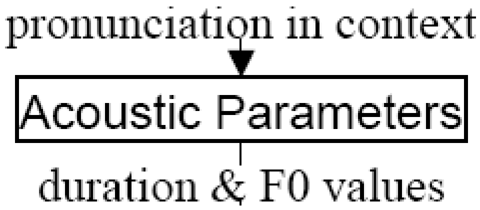
Prosody and meaning

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



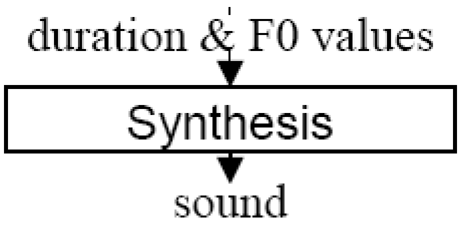
pronunciation in context

Acoustic Parameters

duration & F0 values

- timing:
  - segment duration predicted
    - by rules
    - or 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 🔊

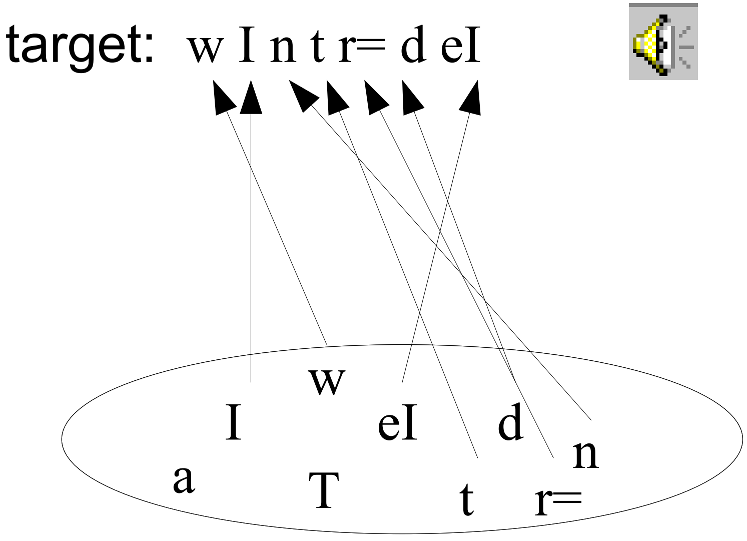
Infovox 🔊

- **HMM-based systems:**

MARY 🔊 🔊

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

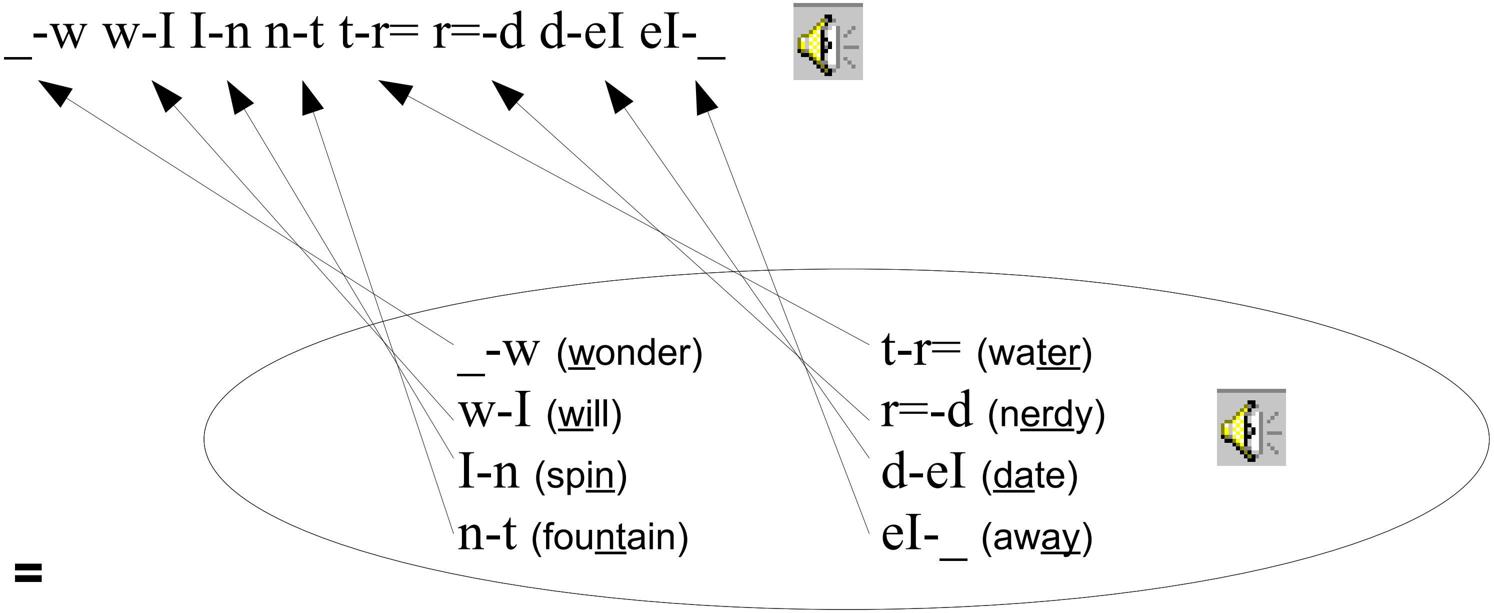
Concatenative synthesis: Isolated phones don't work



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

Concatenative synthesis: Diphones

target: w I n t r= d eI



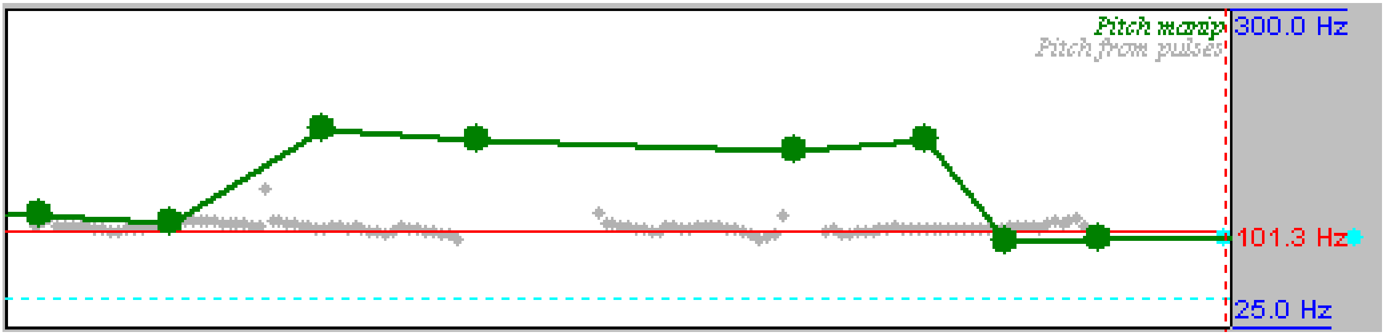
Diphones =
sound segments
from the middle of one phone
to the middle of the next phone

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

Concatenative synthesis: Diphones (2)

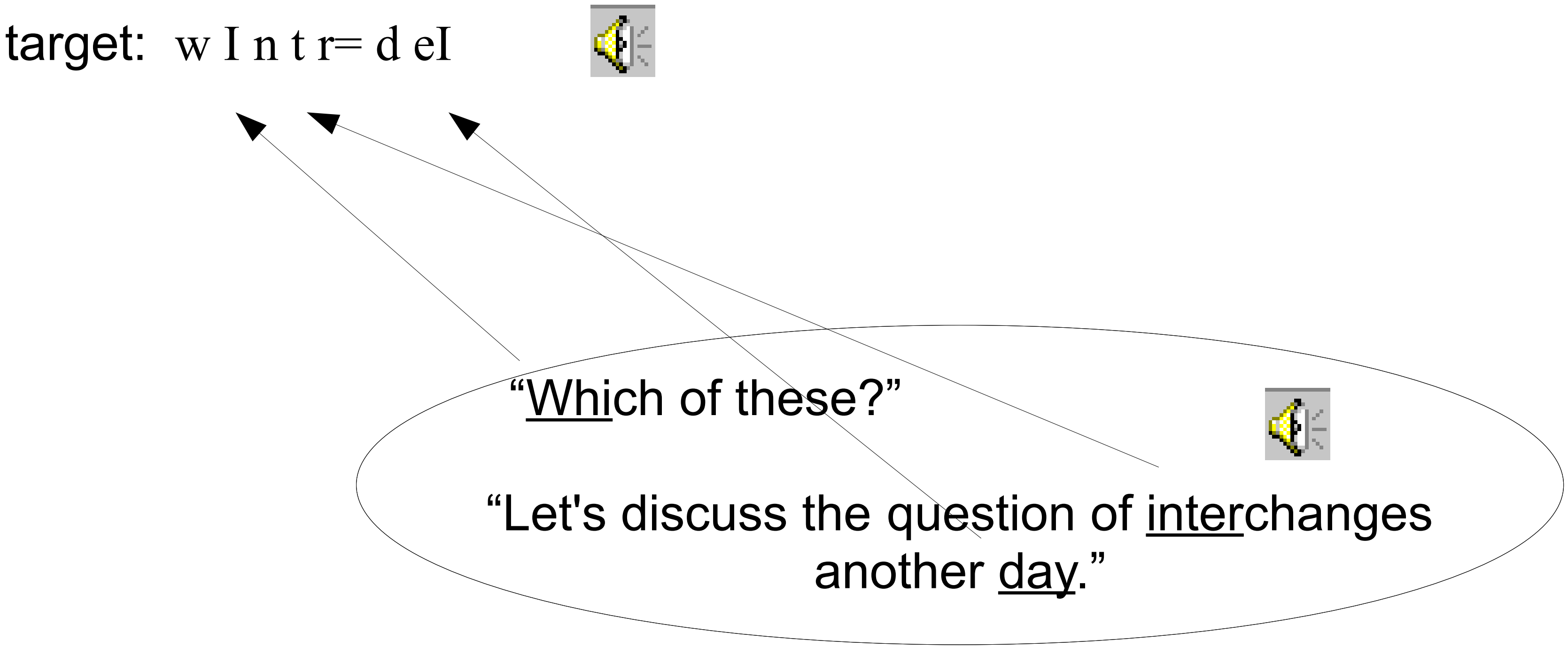
target: w I n t r= d eI

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



PSOLA pitch manipulation

Concatenative synthesis
Unit selection



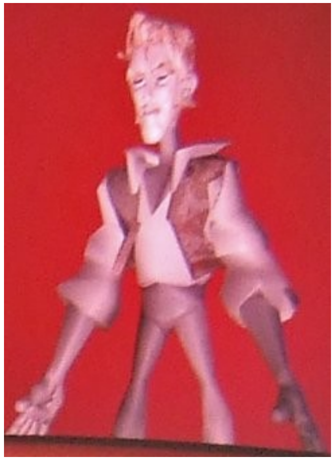
target: w I n t r= d eI

“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)

AI Poker: The voices of Sam and Max



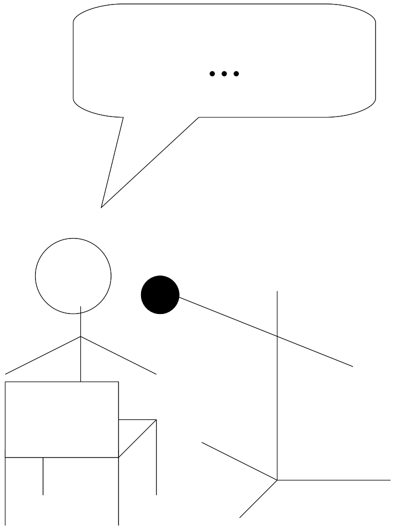
Sam:

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

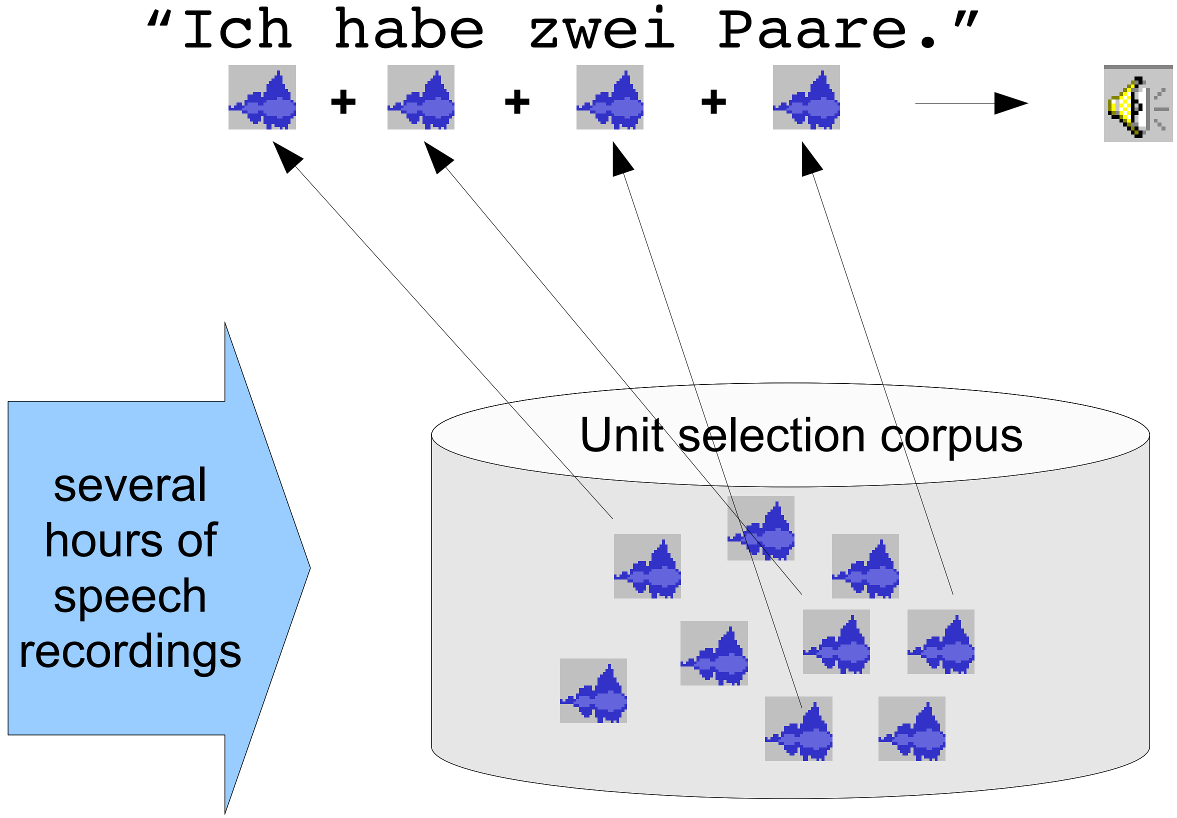
Max:

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

Sam's voice: Unit selection syntheis



…



“Ich habe zwei Paare.”

several hours of speech recordings

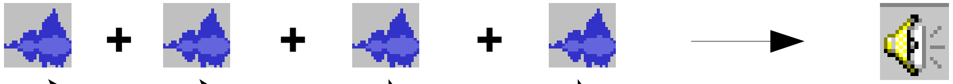
Unit selection corpus

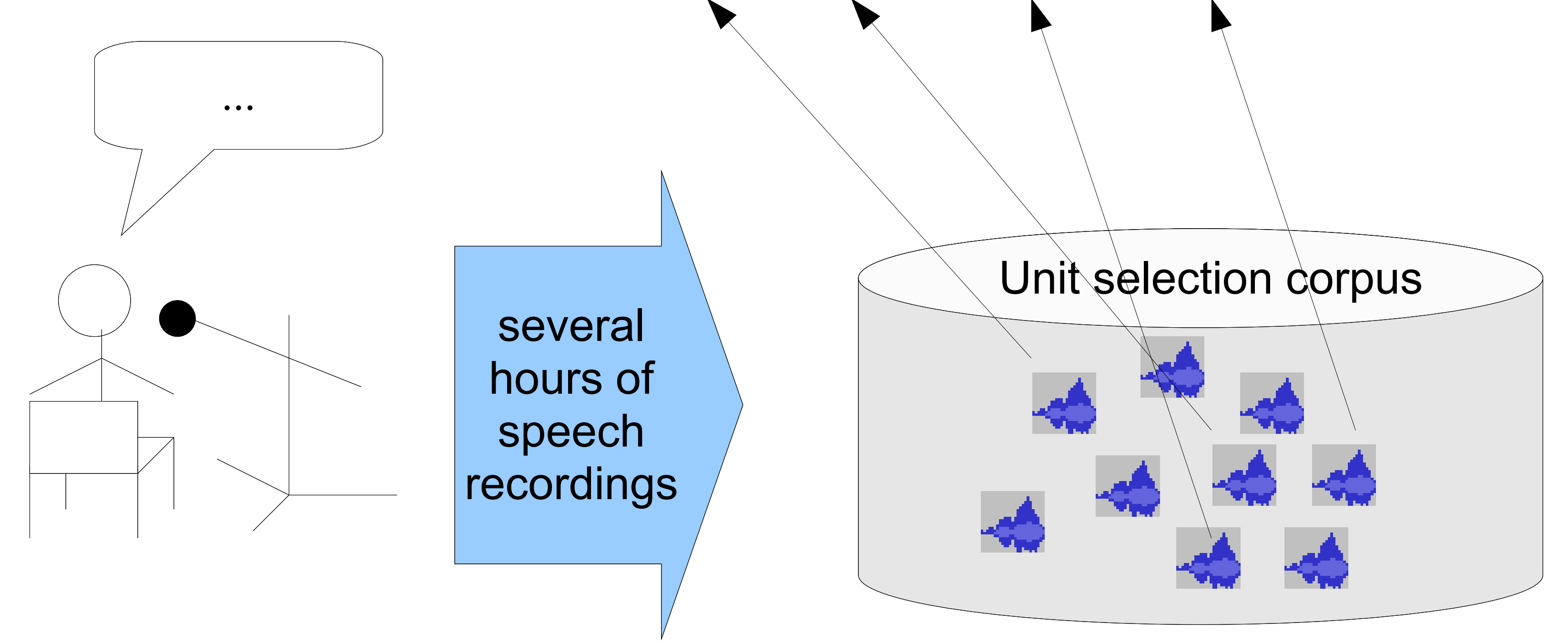
=> very good quality within the poker domain!

Marc Schröder, DFKI

Sam's voice: Unit selection syntheis

“Ich kann auch ganz andere Sachen...”

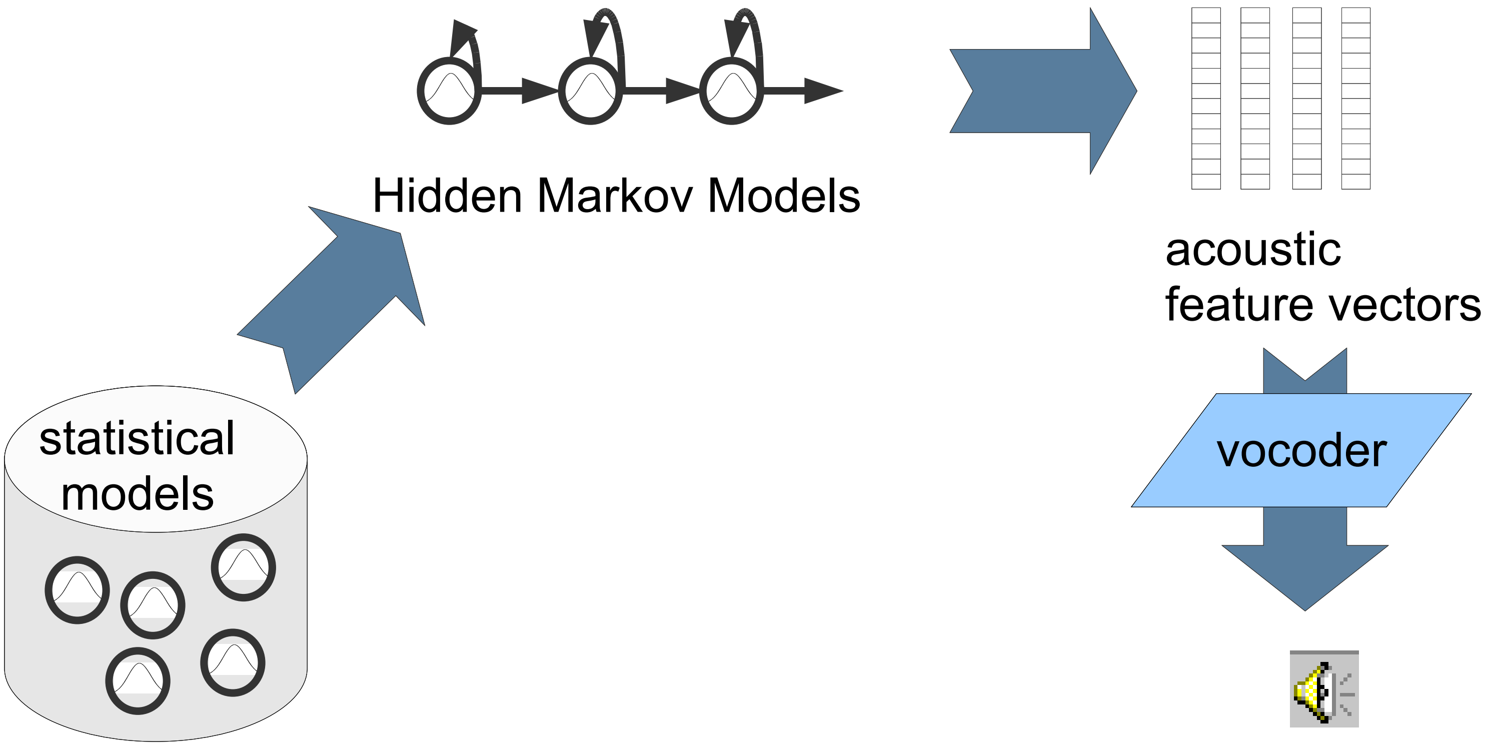




reduced quality with arbitrary text

Max's voice: HMM-based synthesis

“Ich habe zwei Paare.”



Hidden Markov Models

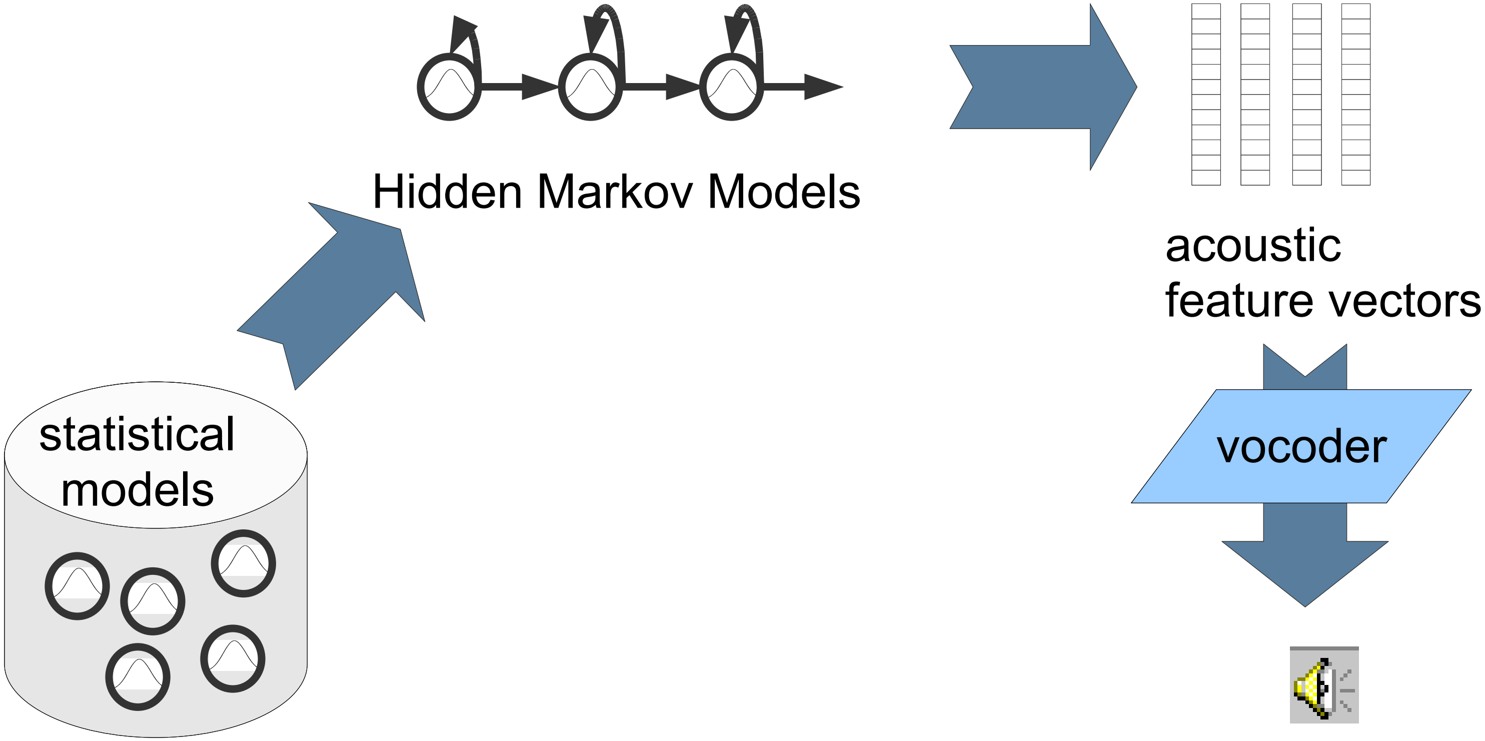
acoustic
feature vectors

vocoder

statistical
models

Max's voice: HMM-based synthesis

“Ich kann auch ganz andere Sachen...”



Hidden Markov Models

acoustic
feature vectors

statistical
models

vocoder

constant quality with arbitrary text

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):		Felix Burkhardt (2001):	
			neutral
	angry		angry
	happy		happy
	sad		sad
	fearful		fearful

Expressive speech synthesis
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:

Marc Schröder (1999): neutral, angry, happy, sad, fearful

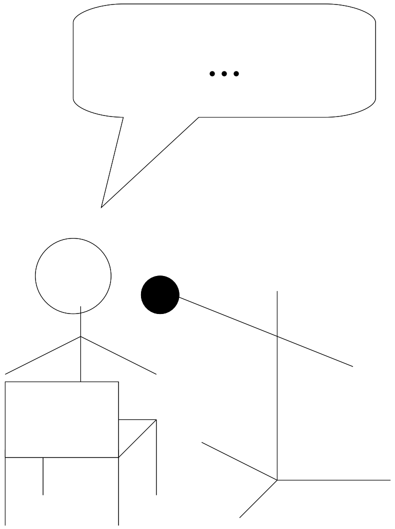
Ignasi Iriondo (2004): angry, happy, sad, fearful

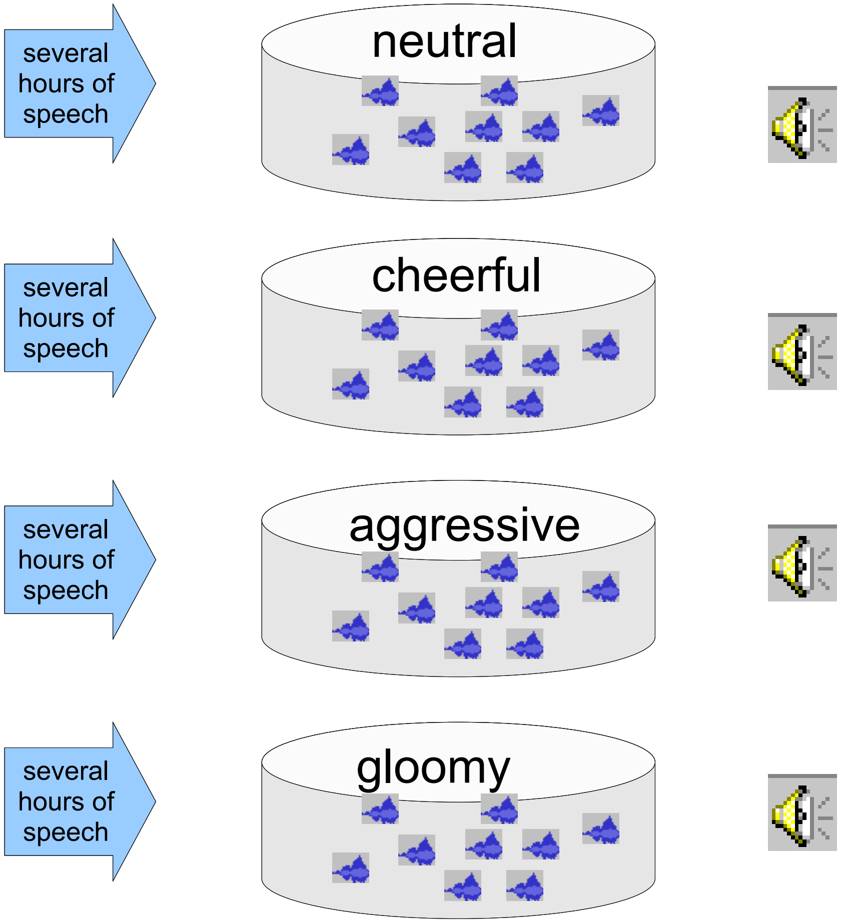
Expressive speech synthesis
Diphone synthesis

- 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)

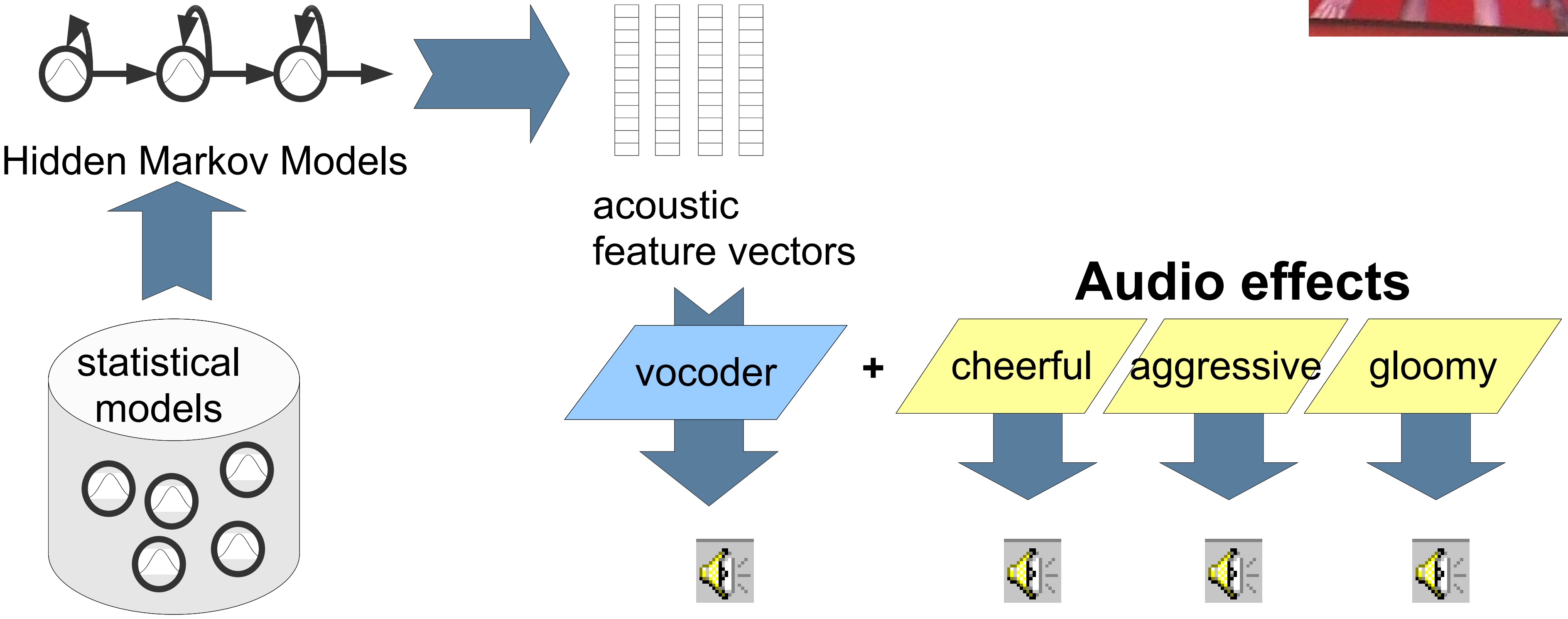
angry1 orig_angry1 angry2 orig_angry2

Sam and the emotions: Expressive unit selection synthesis





Max and the emotions:
Expressive HMM-based synthesis

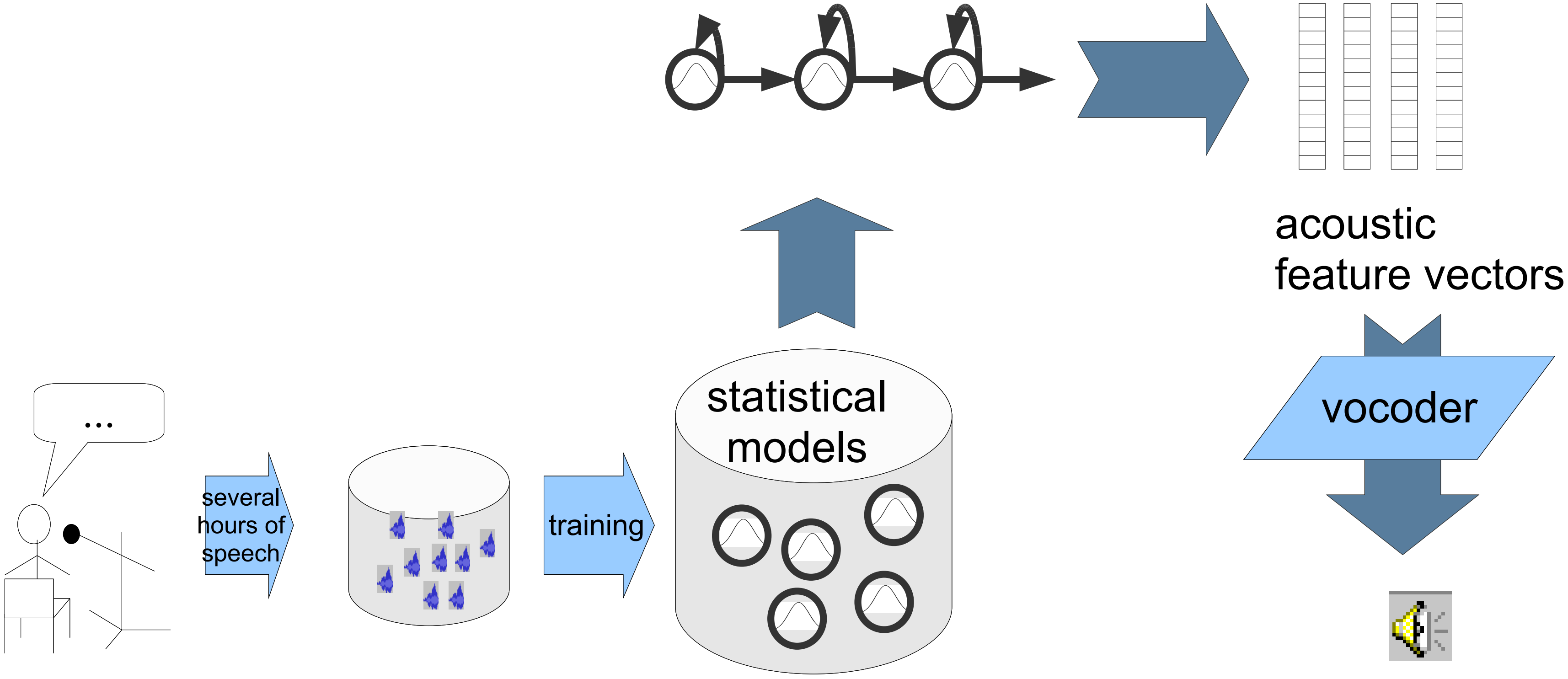


HMM-based synthesis is also data-driven!

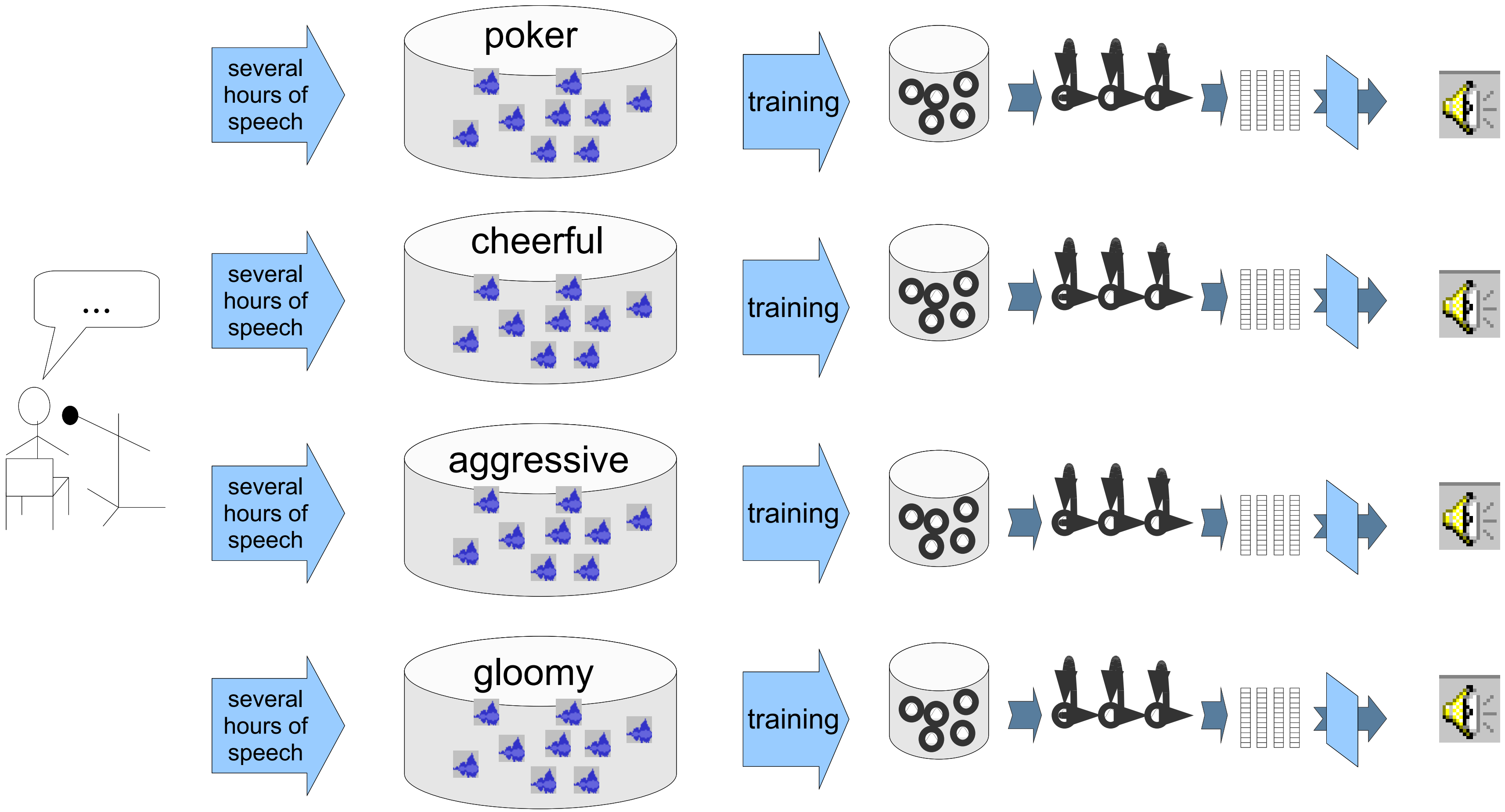
- so far, we have treated the statistical models as given
- thus, expressivity could only be coarsely mimicked using audio effects

... but where do the statistical models come from?!

Statistical models are trained from data



Data-driven expressive HMM-based synthesis



Technologies for expressive TTS: Summary

- “Explicit modelling” approaches
  - low naturalness
  - high flexibility, high control over acoustic parameters
  - explicit models of emotional prosody
- Data-driven approaches
  - expressivity determined by recordings
  - unit selection:
    - high but fragile naturalness, depends on coverage
    - no flexibility, no control over acoustic parameters
  - HMM-based synthesis:
    - medium but constant naturalness
    - some control over acoustic parameters