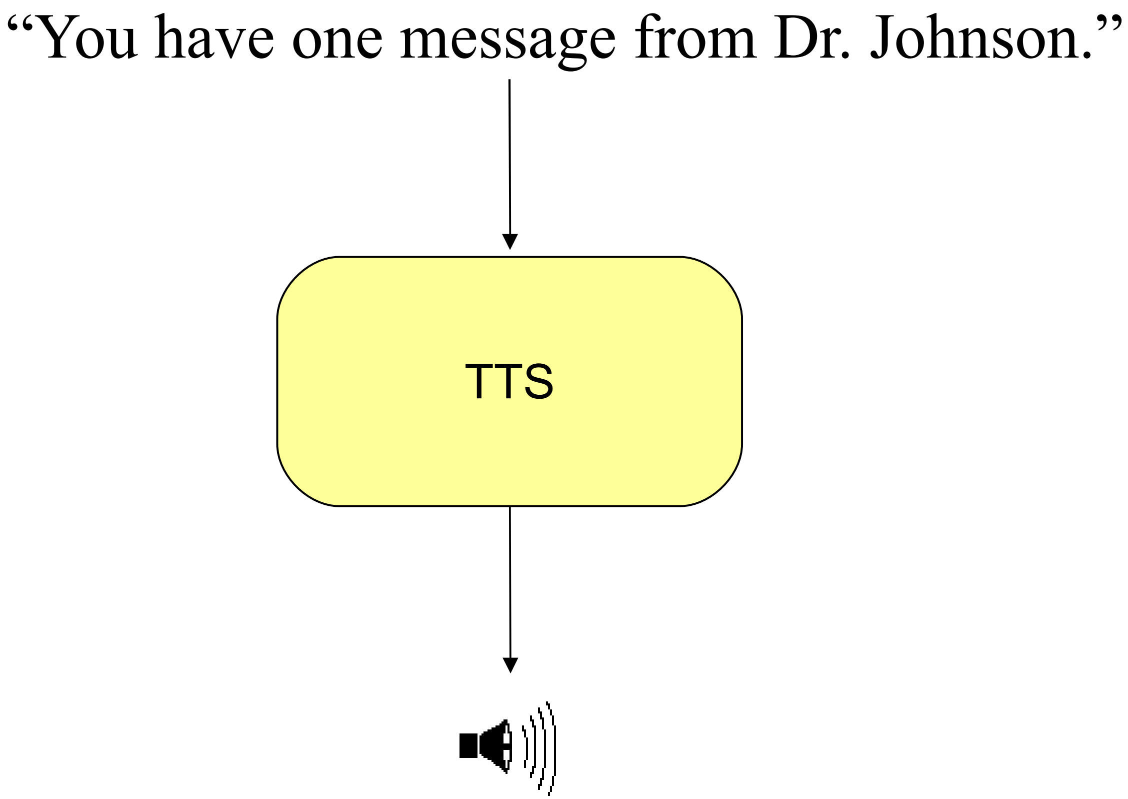
Foundations of Language Science and Technology
Speech synthesis

Marc Schröder, DFKI
schroed@dfki.de

20 January 2010

What is text-to-speech synthesis?



“You have one message from Dr. Johnson.”

TTS

Applications of TTS

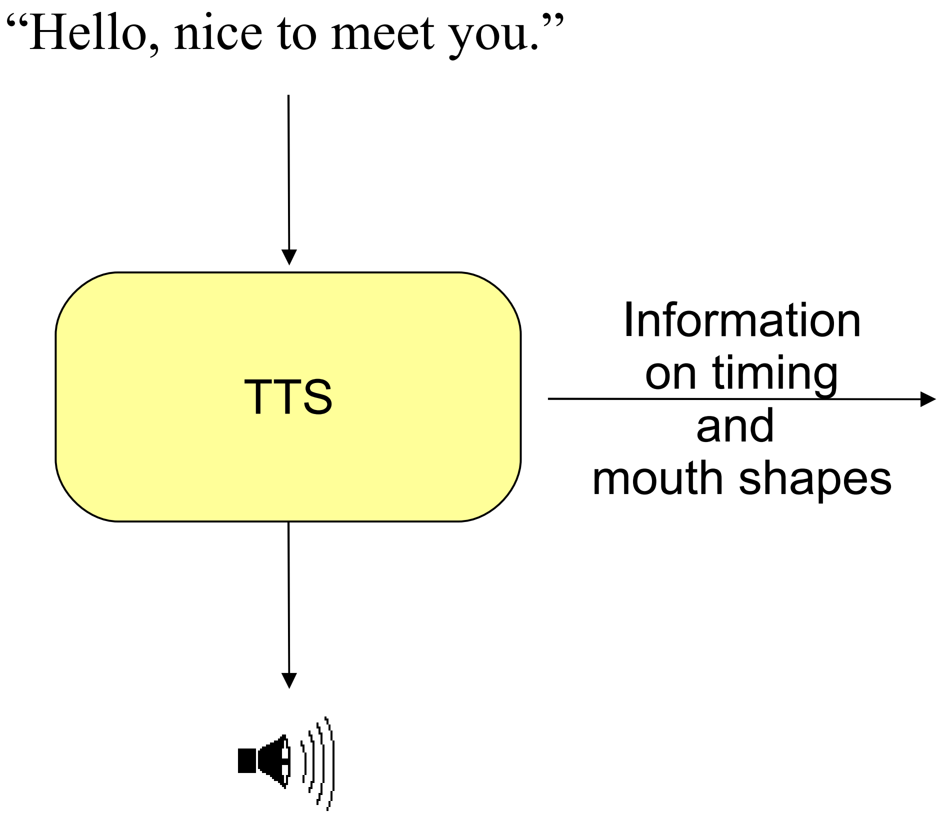
- Texts readers
  - for the blind
  - in eyes-free environments (e.g., while driving)
- Telephone-based voice portals
- Multi-modal interactive systems
  - talking heads
  - “embodied conversational agents” (ECAs)

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



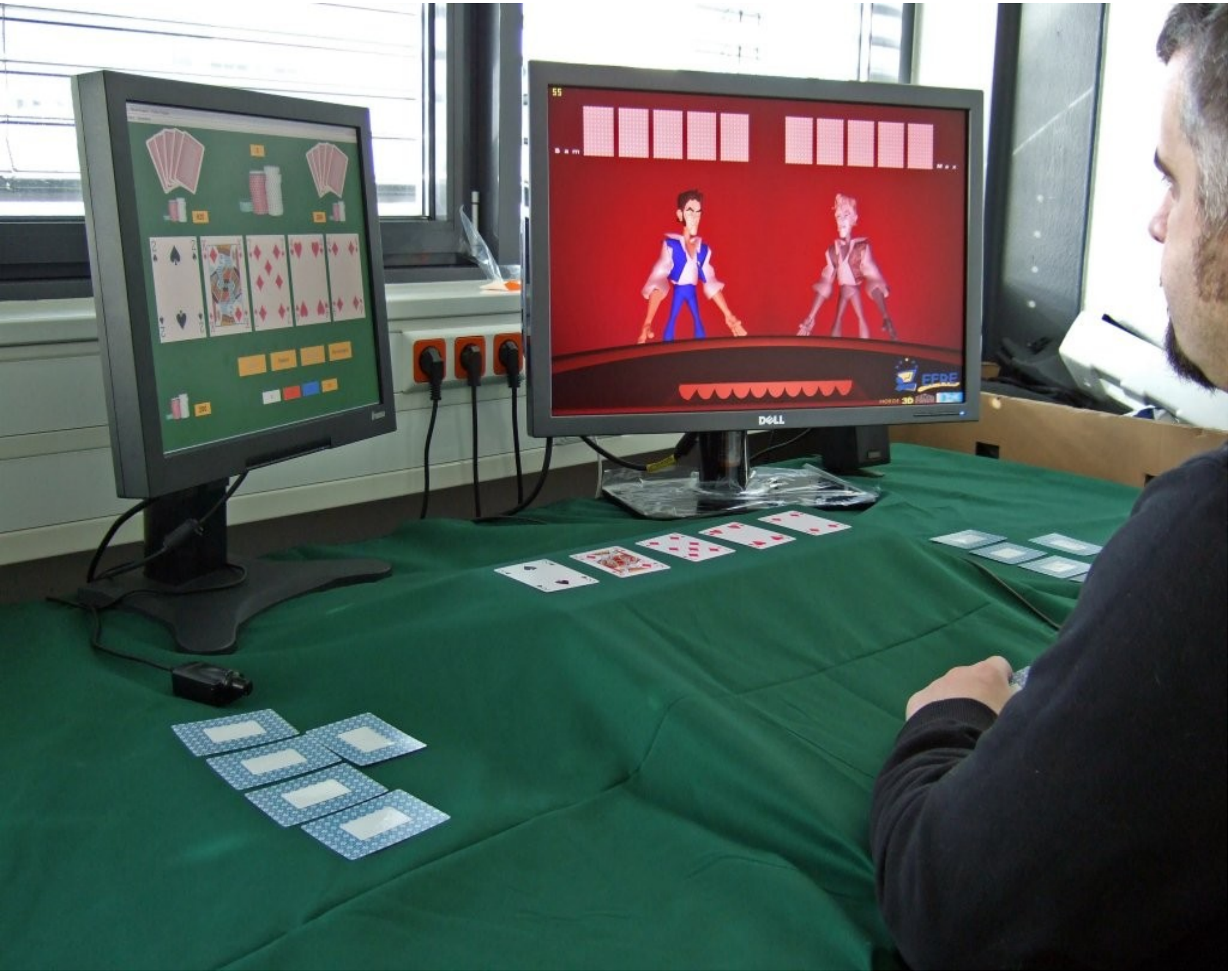
“Hello, nice to meet you.”

TTS

Information on timing and mouth shapes

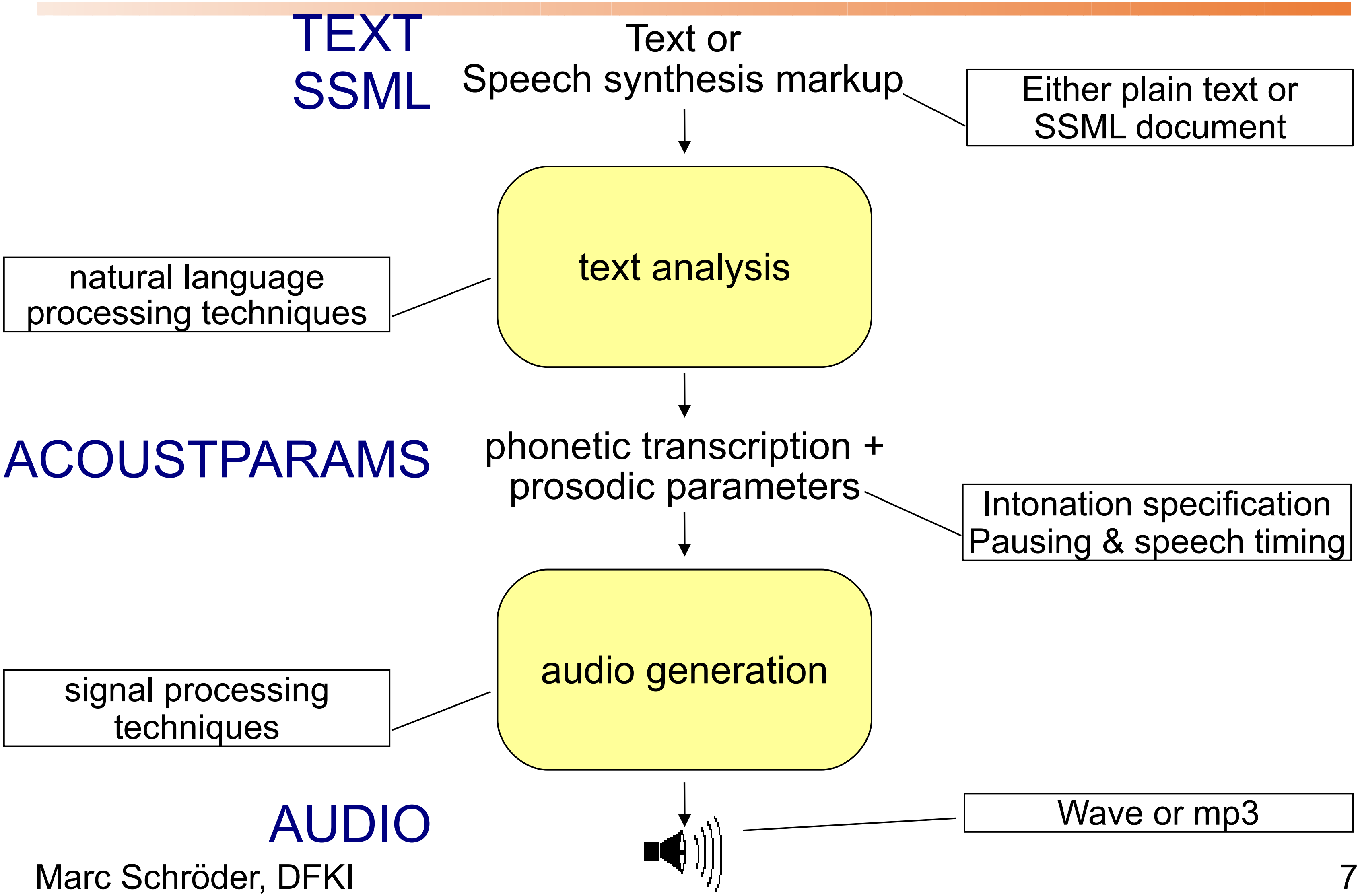
Facial Animation Model,
Computer Graphics Group,
MPI Saarbrücken

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
SSML

Text or
Speech synthesis markup

Either plain text or
SSML document

natural language
processing techniques

text analysis

ACOUSTPARAMS

phonetic transcription +
prosodic parameters

Intonation specification
Pausing & speech timing

signal processing
techniques

audio generation

AUDIO

Wave or mp3

Structure of a TTS system: MARY TTS

- Text analysis
  - Input markup parser TEXT or SSML → RAWMARYXML
  - Shallow NLP RAWMARYXML → PARTSOFSPEECH
  - Phonemiser PARTSOFSPEECH → ALLOPHONES
  - Symbolic prosody ALLOPHONES → INTONATION
  - Acoust. parameters INTONATION → ACOUSTPARAMS
- Audio generation
  - waveform synthesis ACOUSTPARAMS → AUDIO

System structure: Input markup parser

TEXT or SSML → RAWMARYXML

- 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

- Shallow NLP
  - Tokeniser RAWMARYXML → TOKENS
    - sentence boundaries, “tokens” = word-like units
  - Text normalisation TOKENS → WORDS
    - expanded, pronounceable forms (see next slide)
  - Part-of-speech tagger WORDS → PARTSOFSPEECH

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

- Phonemiser PARTSOFSPEECH → PHONEMES
  - lexicon lookup
  - letter-to-sound conversion
    - morphological decomposition
    - letter-to-sound rules
    - syllabification
    - word stress assignment
- Custom pronounciation PHONEMES → ALLOPHONES
  - slurring, non-standard pronounciation
  - potentially trainable from annotated data of a given person

System structure: Prosody

- “Prosody”?
  - intonation (accented syllables; high or low phrase boundaries)
  - rhythmic effects (pauses, syllable durations)
  - loudness, voice quality
- Symbolic prosody prediction

ALLOPHONES → INTONATION

  - 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

- Duration prediction INTONATION → DURATIONS
  - segment duration predicted
    - by rules
    - or by decision trees
- Contour generation DURATIONS → ACOUSTPARAMS
  - fundamental frequency curve predicted
    - by rules
    - or by decision trees

System structure: Waveform synthesis

- Waveform synthesis ACOUSTPARAMS → AUDIO
  - several waveform generation technologies

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 🔊

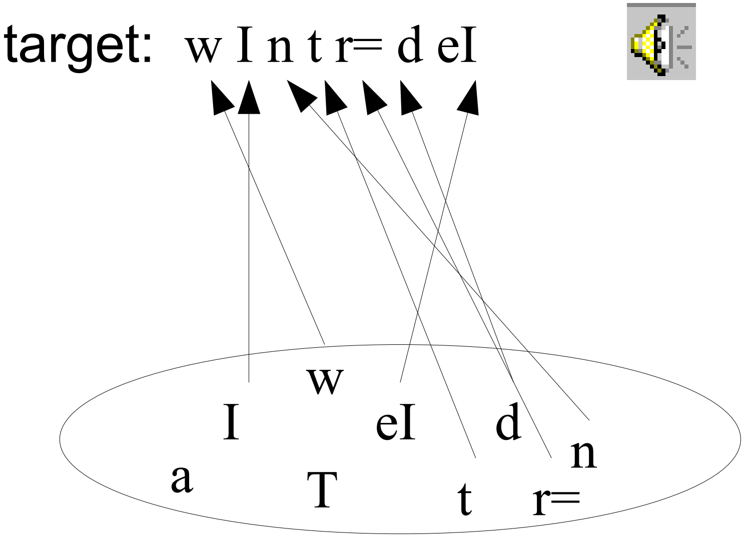
Infovox 🔊

- **HMM-based systems:**

MARY 🔊 🔊

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

Concatenative synthesis:
Isolated phones don't work



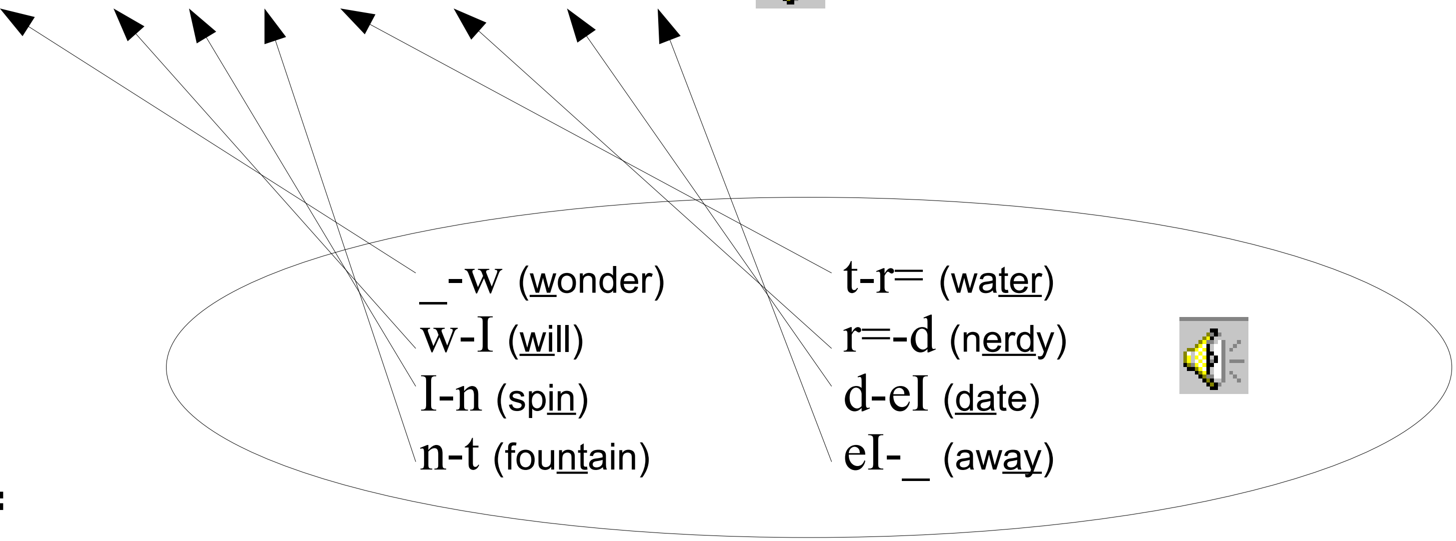
target: w I n t r= d eI

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)

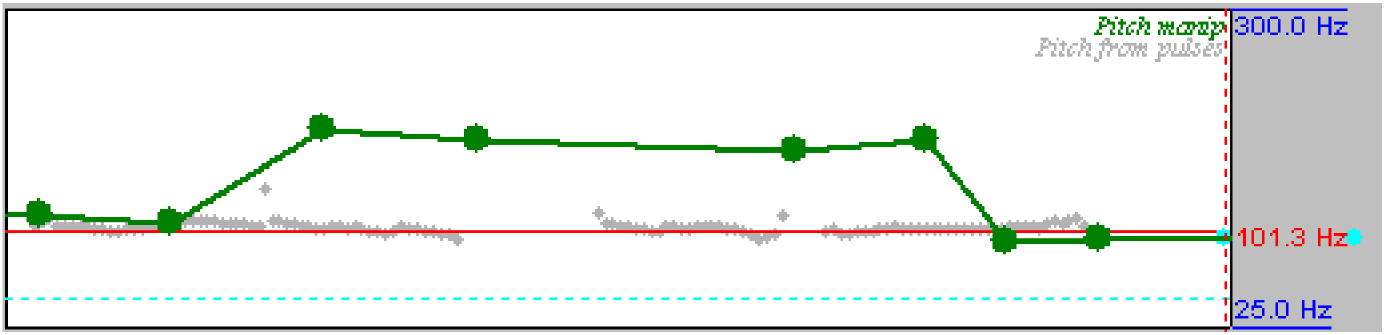
Diphones =
sound segments
from the middle of one phone
to the middle of the next phone

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

Concatenative synthesis: Diphones (2)

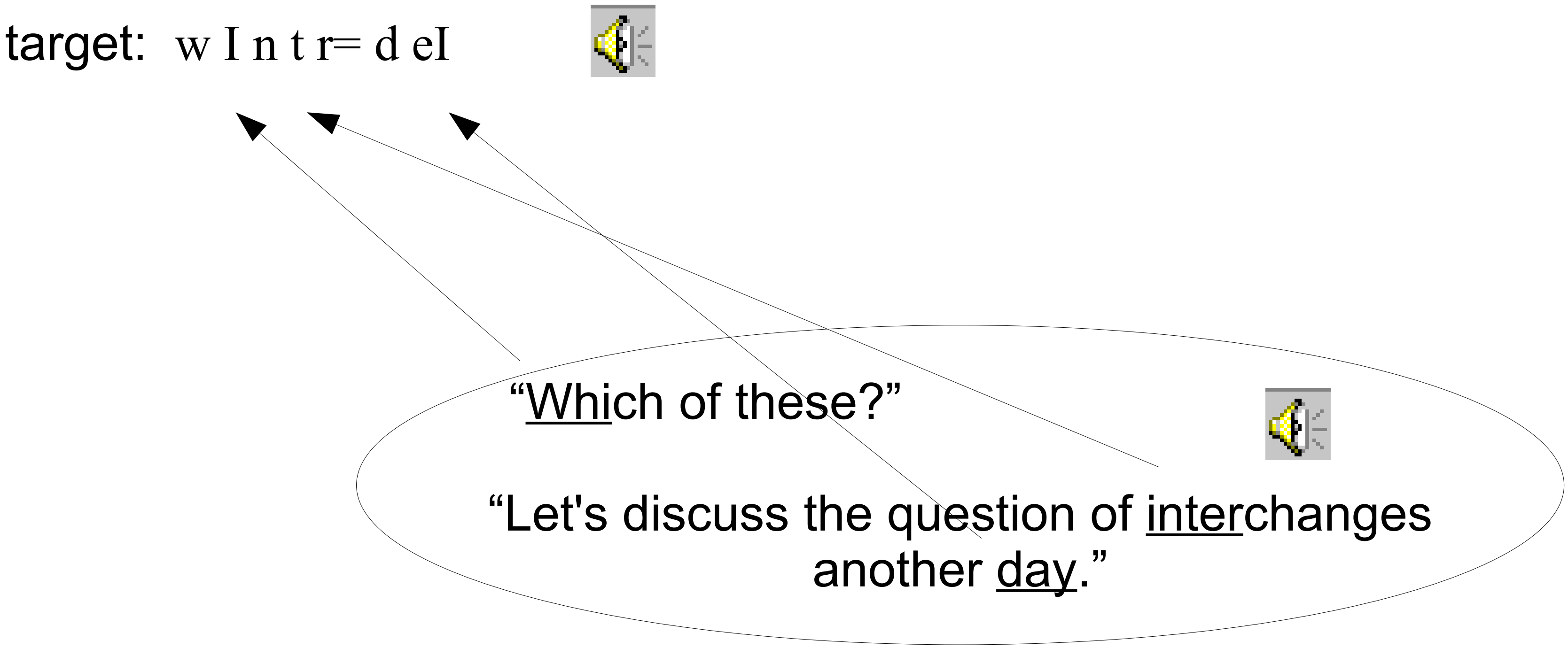
target: w I n t r= d eI

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



PSOLA
pitch
manipulation

Concatenative synthesis
Unit selection



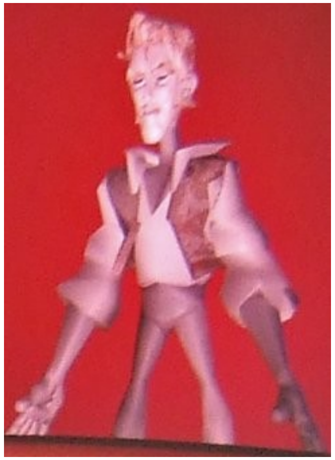
target: w I n t r= d eI

“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)

AI Poker: The voices of Sam and Max



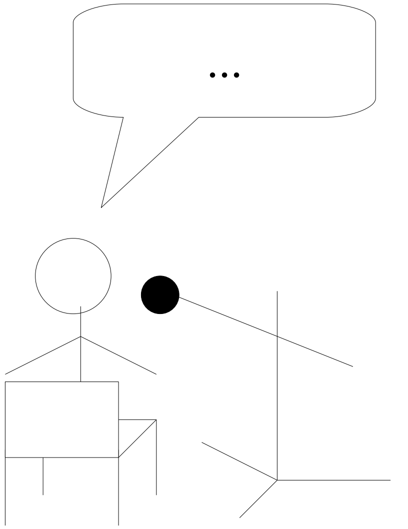
Sam:

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

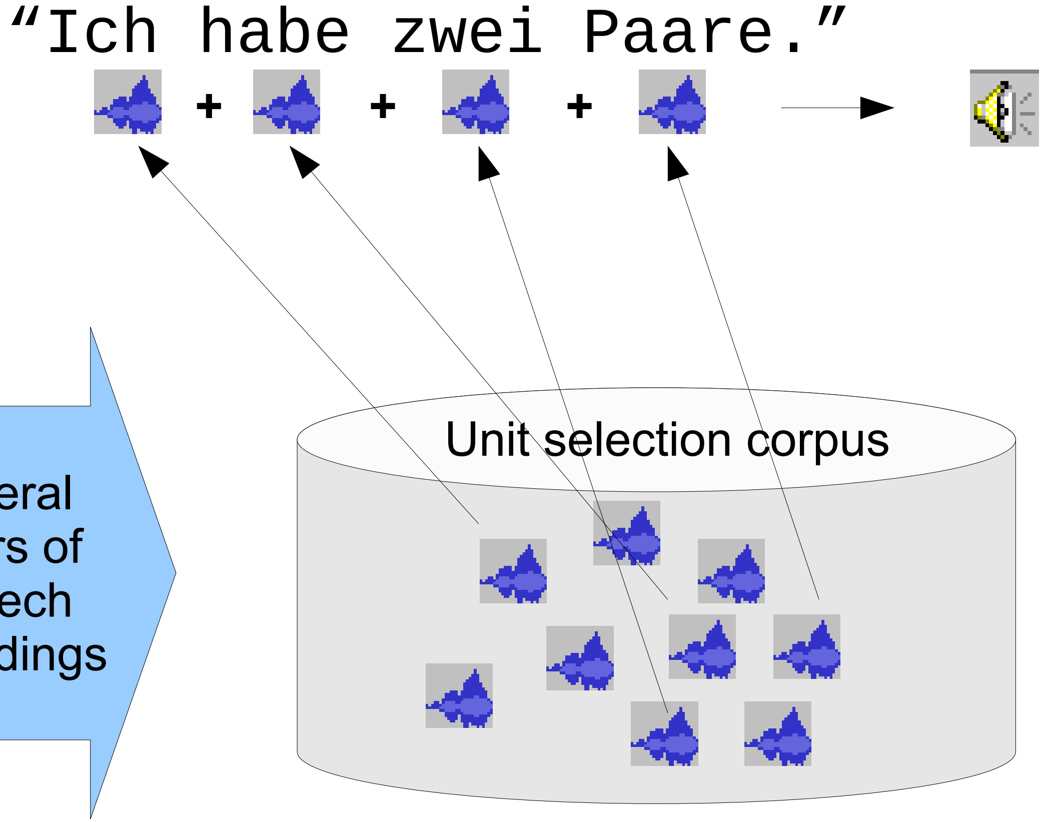
Max:

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

Sam's voice: Unit selection syntheis



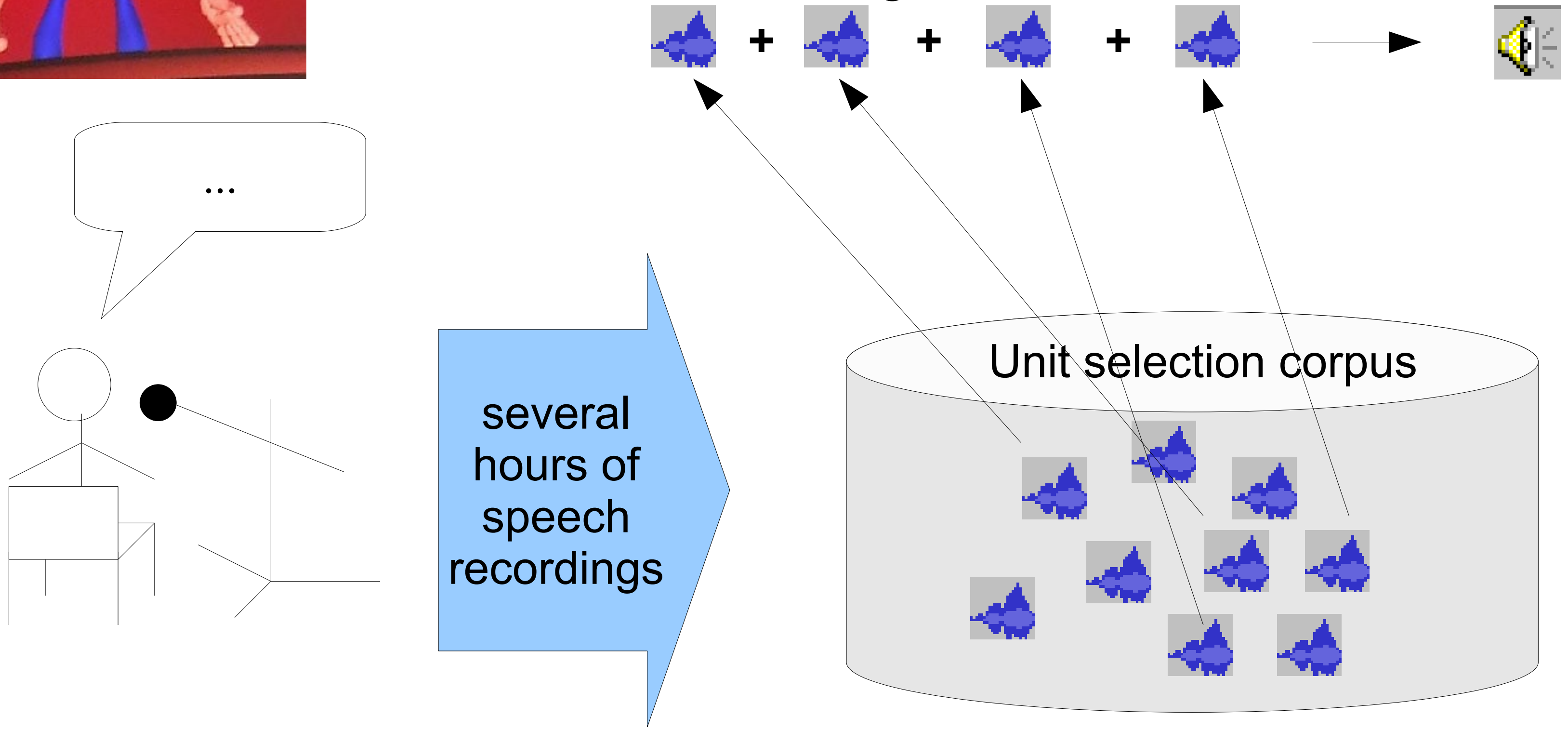
several hours of speech recordings


“Ich habe zwei Paare.”

=> very good quality within the poker domain!

Sam's voice: Unit selection syntheis

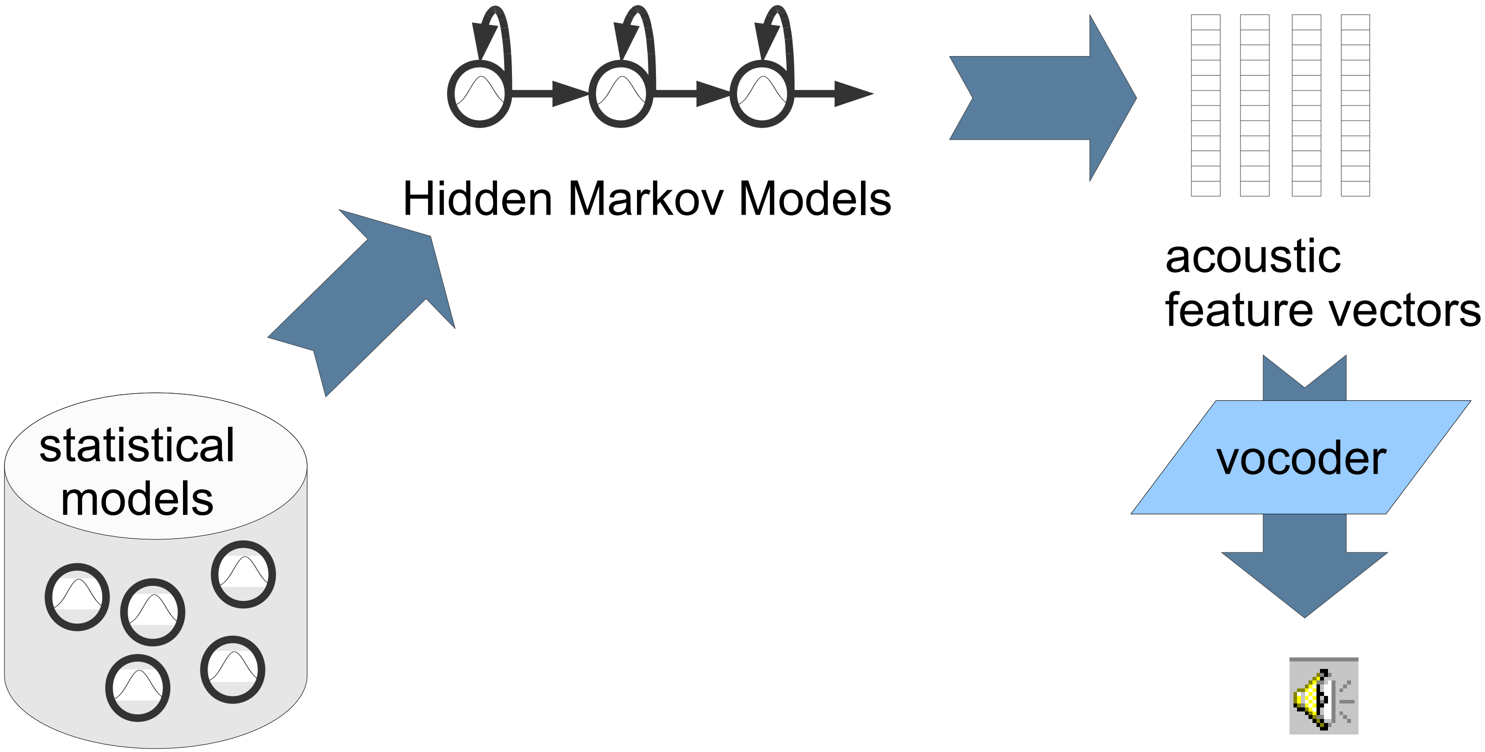
“Ich kann auch ganz andere Sachen...”



reduced quality with arbitrary text

Max's voice: HMM-based synthesis

“Ich habe zwei Paare.”


Hidden Markov Models

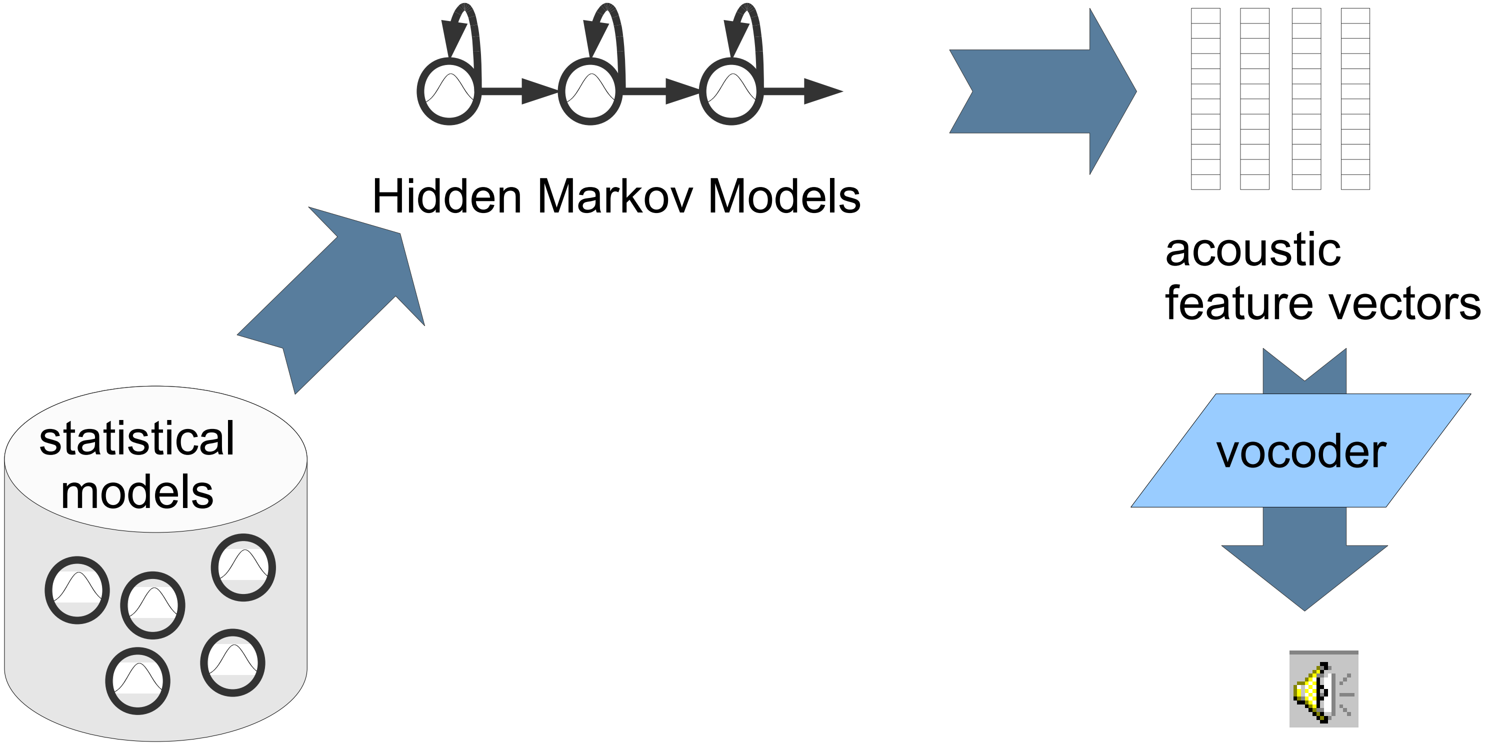
acoustic
feature vectors

vocoder

statistical
models

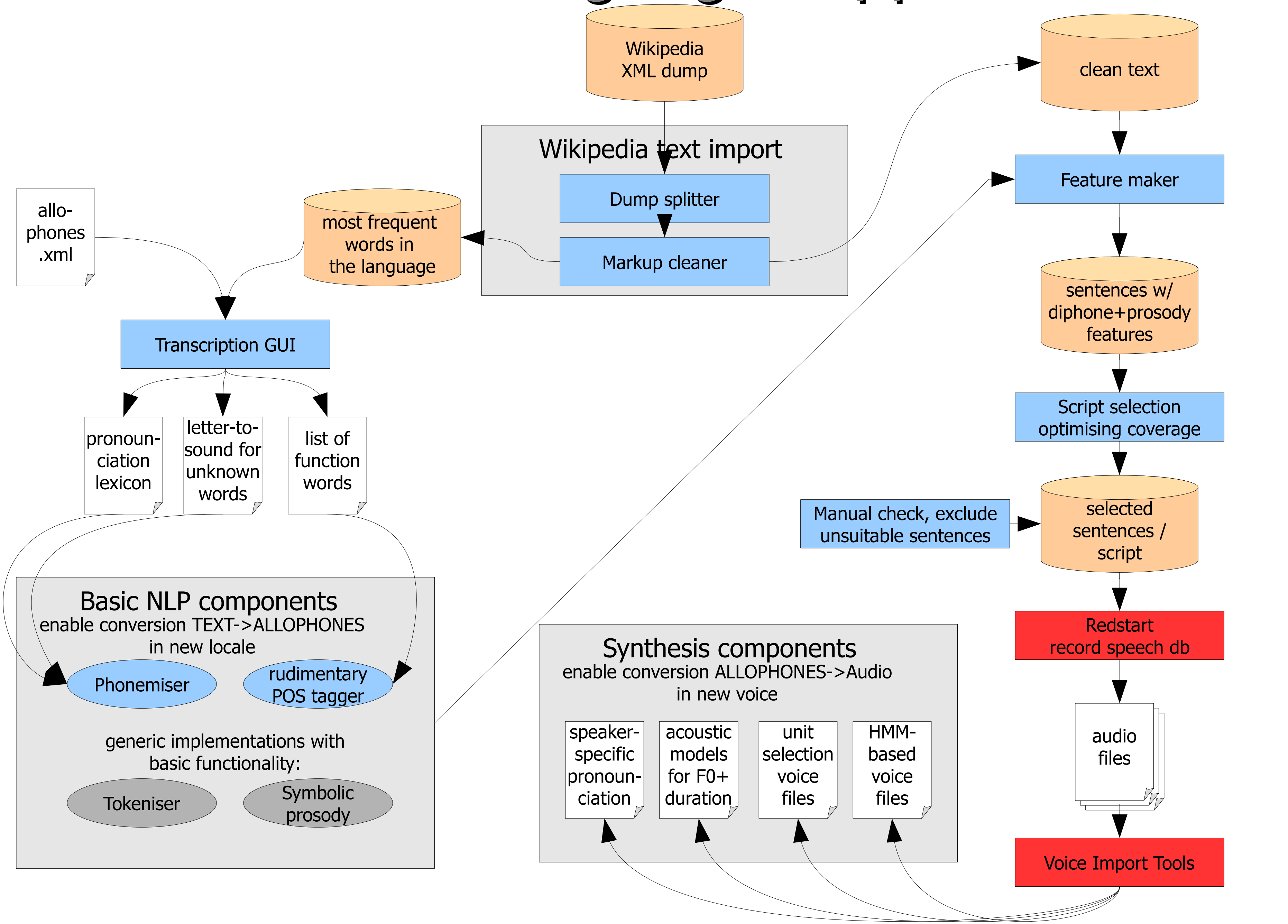
Max's voice: HMM-based synthesis

“Ich kann auch ganz andere Sachen...”



constant quality with arbitrary text

MARY TTS: New language support workflow



Wikipedia XML dump

Wikipedia text import

Dump splitter

Markup cleaner

most frequent words in the language

allo-phones .xml

Transcription GUI

pronoun-ciation lexicon

letter-to-sound for unknown words

list of function words

Basic NLP components

enable conversion TEXT->ALLOPHONES in new locale

Phonemiser

rudimentary POS tagger

generic implementations with basic functionality:

Tokeniser

Symbolic prosody

clean text

Feature maker

sentences w/ diphone+prosody features

Script selection optimising coverage

Manual check, exclude unsuitable sentences

selected sentences / script

Redstart record speech db

audio files

Voice Import Tools

Synthesis components

enable conversion ALLOPHONES->Audio in new voice

speaker-specific pronoun-ciation

acoustic models for F0+ duration

unit selection voice files

HMM-based voice files

Hands-on TTS: MARY TTS 4.0

- Get it from http://mary.dfki.de
  - either download onto your machine (~32 MB min download)
  - or use online demo