Specialization Module

Speech Technology

Timo Baumann
baumann@informatik.uni-hamburg.de

UNIVERSITÄT HAMBURG, DEPARTMENT OF INFORMATICS
NATURAL LANGUAGE SYSTEMS GROUP

not „just“ Text-to-Speech Synthesis

Synthesis examples

- first singing (digital) computer (IBM, 1961)
  → hand-tuned vocoding
- extension of the same technique today: espeak
  → rule-based vocoding system
- based on natural speech: DreSS-FR, Mbrola
  → diphone-synthesis
- a more modern system: MaryTTS
  → general concatenative speech synthesis
- smaller memory footprint of the above
  → HMM-based speech synthesis (to be covered in 2 weeks)

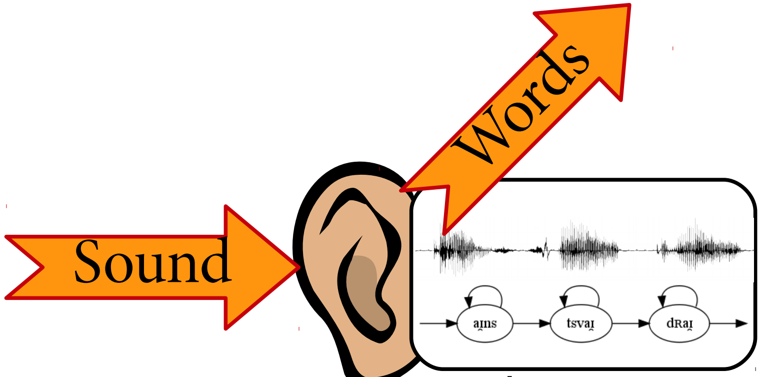
Input and Output for Spoken Dialogue Systems

- Recognition
  - Reduction of the signal to words

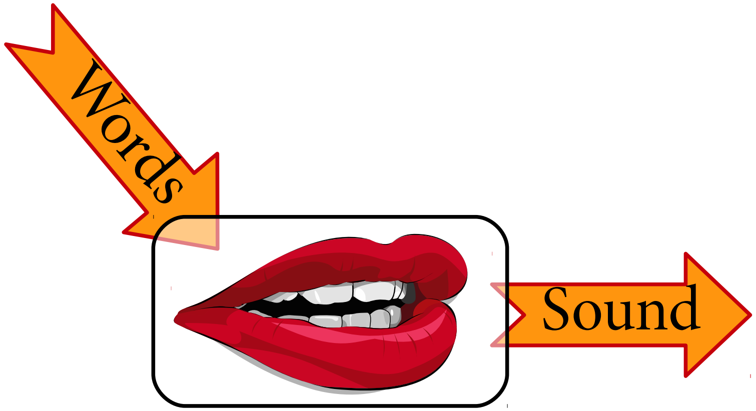
  ➔ *abstraction* from details

- Synthesis
  - ▪ words themselves only insufficiently describe the signal

  ➔ naturalness only with *addition* of details



Speech Recognition



Speech Synthesis

what is *missing* in written language?

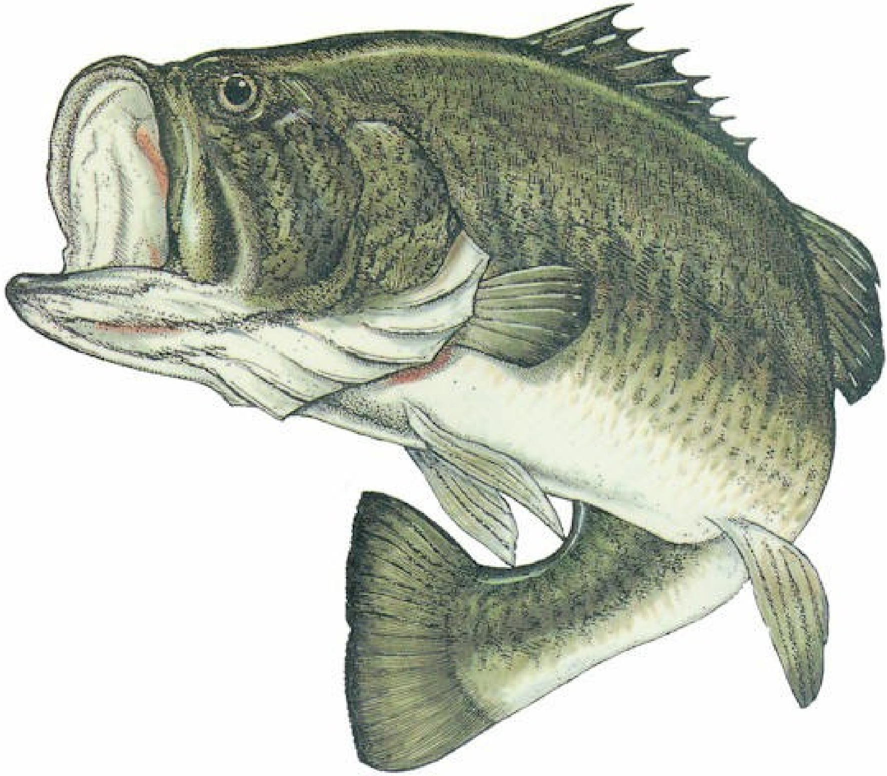
Written vs. Spoken Language
Timo's list

- Abbreviations, dates, numbers, currencies, …
- Homographs: Bass
- Text does not have any melody or rhythm!
  - prosody is important to convey meaning
  - Punctuation only partially helpful

Homographs

[baɪs] [bæs]

Bass



information structure

Information Structure

The linguistic means of structuring information, in order to optimize information transfer within discourse

  - Topic / Focus
  - Given / New information

- not directly conveyed in textual representation
  - but to a certain degree by prosody
- to reconstruct the structure, listeners also use
  - context of the utterance in the whole conversation
  - world knowledge

Sonderforschungsbereich 2003-2015: http://www.sfb632.uni-potsdam.de

Prosody

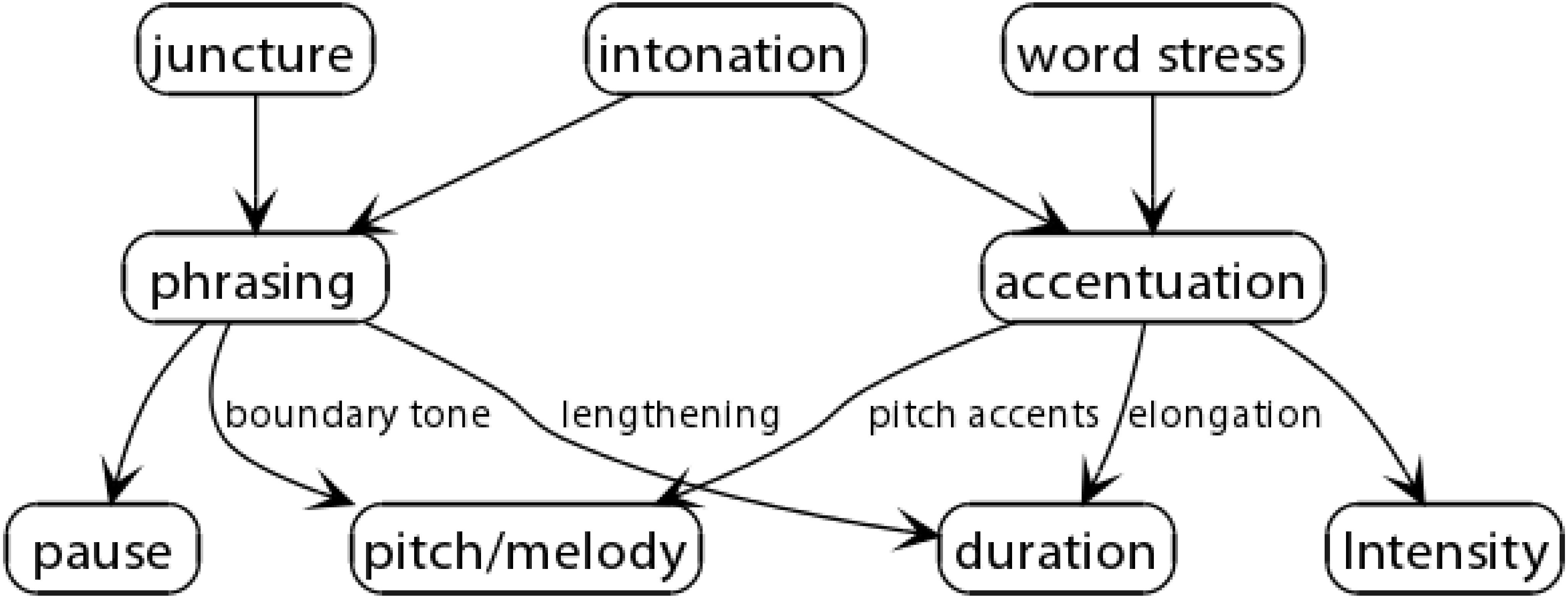
supra-segmental properties of speech

- phenomena:
  - pitch (i.e., melody / fundamental frequency)
  - loudness / intensity
  - duration, pauses

- phonetically: accentuation and phrasing

- phonologically: (word)stress, intonation, juncture

Prosody:
Phonology – Phonetics – Phenomena



Focus and Accentuation

Focus and Accentuation

- “I didn't say we should kill him.”
  - someone else said we should kill him
  - I am denying that I said we should kill him
  - I wrote it down or implied it, but I didn't say it
  - I said someone else should do the job
  - I said that we absolutely must kill him
  - getting him a little nervous would have been enough
  - we got the wrong guy

Focus and Accentuation

- “I didn't say we should kill him.”
  - someone else said we should kill him
  - I am denying that I said we should kill him
  - I wrote it down or implied it, but I didn't say it
  - I said someone else should do the job
  - I said that we absolutely must kill him
  - getting him a little nervous would have been enough
  - we got the wrong guy

Focus and Accentuation

- “I didn't say we should kill him.”
  - someone else said we should kill him
  - I am denying that I said we should kill him
  - I wrote it down or implied it, but I didn't say it
  - I said someone else should do the job
  - I said that we absolutely must kill him
  - getting him a little nervous would have been enough
  - we got the wrong guy

Focus and Accentuation

- “I didn't say we should kill him.”
  - someone else said we should kill him
  - I am denying that I said we should kill him
  - I wrote it down or implied it, but I didn't say it
  - I said someone else should do the job
  - I said that we absolutely must kill him
  - getting him a little nervous would have been enough
  - we got the wrong guy

Focus and Accentuation

- “I didn't say we should kill him.”
  - someone else said we should kill him
  - I am denying that I said we should kill him
  - I wrote it down or implied it, but I didn't say it
  - I said someone else should do the job
  - I said that we absolutely must kill him
  - getting him a little nervous would have been enough
  - we got the wrong guy

Focus and Accentuation

- “I didn't say we should kill him.”
  - someone else said we should kill him
  - I am denying that I said we should kill him
  - I wrote it down or implied it, but I didn't say it
  - I said someone else should do the job
  - I said that we absolutely must kill him
  - getting him a little nervous would have been enough
  - we got the wrong guy

Focus and Accentuation

- “I didn't say we should kill him.”
  - someone else said we should kill him
  - I am denying that I said we should kill him
  - I wrote it down or implied it, but I didn't say it
  - I said someone else should do the job
  - I said that we absolutely must kill him
  - getting him a little nervous would have been enough
  - we got the wrong guy

Information Structure

- information structure is an active area of research:
  - unknown how exactly to represent IS (cross-linguistically, cross-genre, in dialogue, …)
  - unknown how (exactly) IS influences speech

- problem of premature implementation:

**can we really expect a computer
to successfully perform speech synthesis
even before the basic research has been done?**

What a computer *can* do

- problems that are well understood:
  - find solutions based on a model
  - use lists of exceptions if model is faulty
- problems that are somewhat understood:
  - use heuristics to get details right
  - try to avoid taking a stand
- problems that aren't yet understood:
  - require additional instructions in the input
  - guess

What a computer *can* do: focus

- human listeners are predictive (and forgiving):
  - it's worse to be very wrong occasionally
    than to say everything a little bit wrongly
  - human listeners will select the correct interpretation
    (using *their* world knowledge) from available options
- solution:
  - put a small accentuation on all possible focus points
- however
  - system does not *take a stand*, it sounds indifferent, bored

Process diagram of Speech Synthesis

input

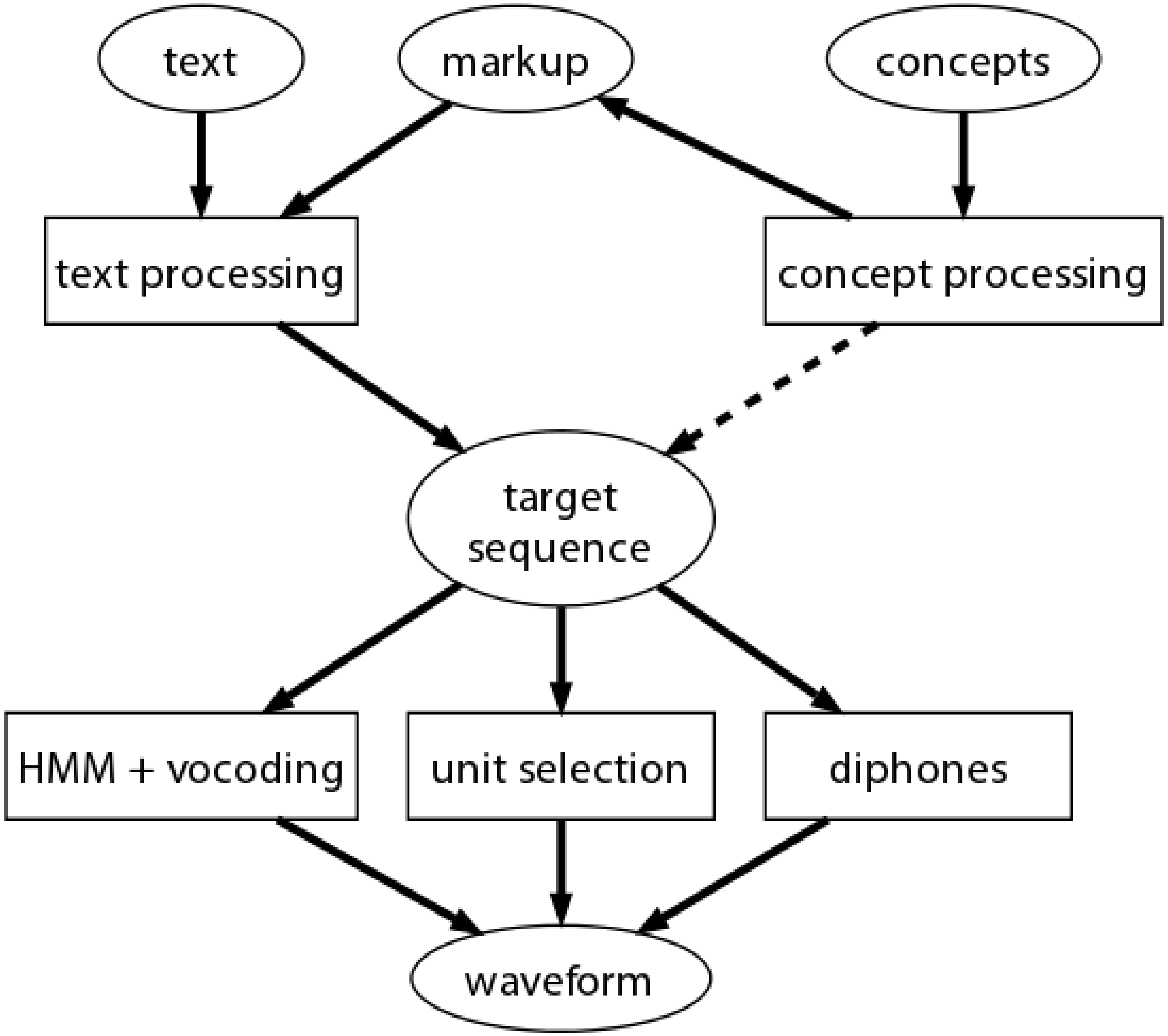
✓ symbolic processing

problem: text is missing detail

intermediate
representation

waveform synthesis

output



Process diagram of Speech Synthesis

input

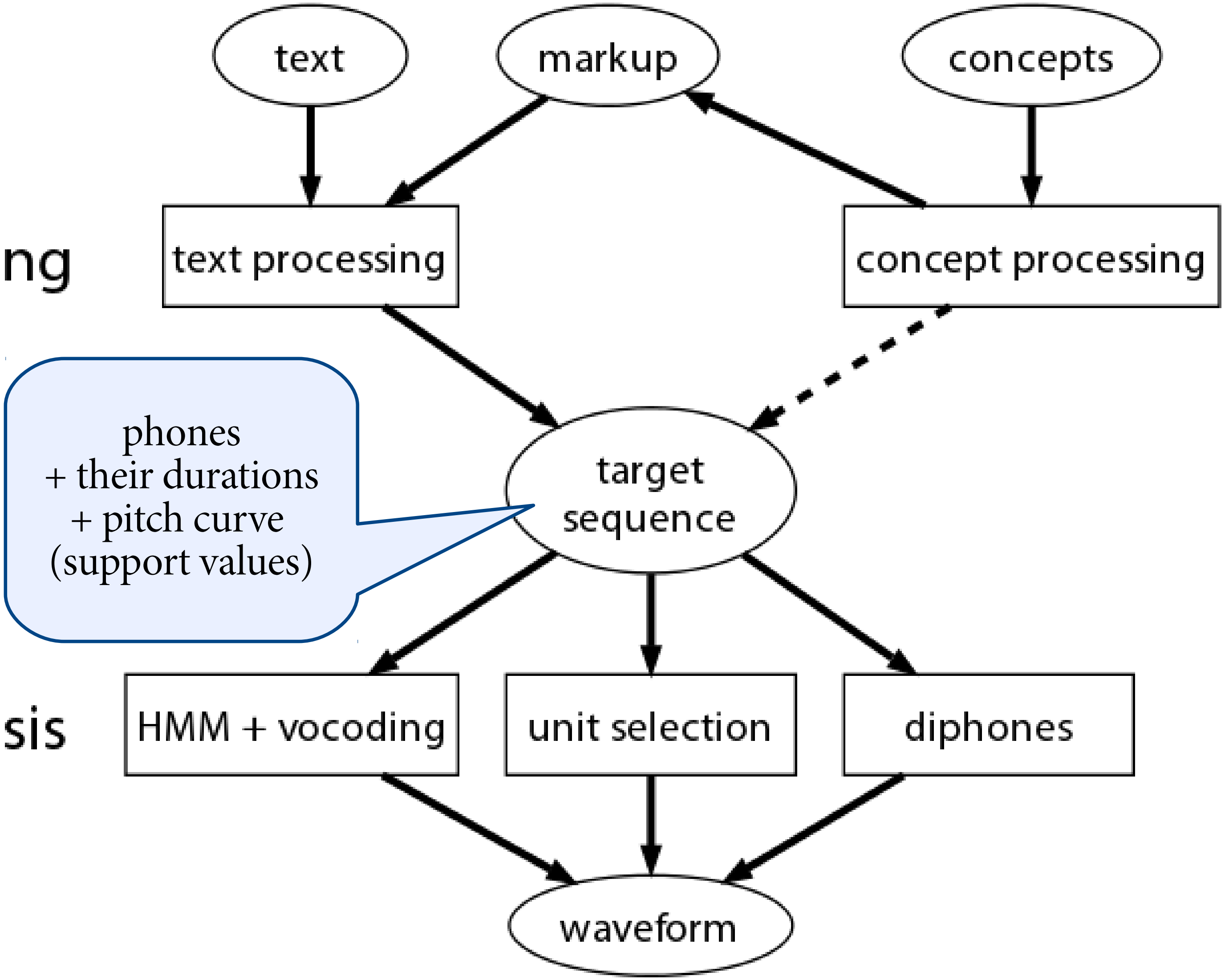
✓ symbolic processing

problem: text is missing detail

intermediate representation

waveform synthesis

output



waveform synthesis

Waveform Synthesis

from the target sequence (phones+duration+pitch)

1. formant-based:

   rules to determine target formants and other parts of the signal
   rules to determine transitions

2. pattern-based:

   database of many short speech segments
   segments are concatenated one after the other

3. model-based approach in 2 weeks

Diphone Synthesis

- Concatenation of short speech snippets
- units from center of a phone to center of the next:
  _h+ha:+a:l+lo:+o:_+_v+vi:+i:g+ge:+e:t+ts+s_
  - concatenation within “stable” phase of the phone
  - coarticulation is (largely) covered
- 40 phones → ~1600 diphones!
  - recorded from one speaker ⌧ one voice
  - additional signal processing for duration+pitch change

General Concatenative Synthesis

- alternatives for the mapping target → speech snippets
  - more speech material in database
  - selection of material that better fits the target sequence
- selection becomes a search of best concatenation
  - costs of fit of concatenation between snippets
  - costs of fit of snippets to target sequence
- computationally expensive (search)
  - very high memory demands (500MB+ per voice)
- results can be very natural sounding

what do you *like* better:
formant-based or pattern-based synthesis?

Summary

Thank you.

baumann@informatik.uni-hamburg.de

https://nats-www.informatik.uni-hamburg.de/SLP16

Further Reading

- Speech Synthesis in General:
  - P. Taylor (2009): *Text-to-Speech Synthesis*. Cambridge Univ Press. ISBN: 978-0521899277. InfBib: A TAY 43070.
- The MaryTTS Speech Synthesis System:
  - Schröder & Trouvain (2003): "The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching", *Int. J. of Speech Technology* **6**(3).

Notizen

Desired Learning Outcomes

- speech synthesis goal is to add variation for naturalness (this is opposite from ASR)
- problems/ambiguities in linguistic pre-processing
  - prosody and pitch: ToBI, information structure
  - major synthesis techniques: formants, diphone,
  - (PSOLA technique)