Multimodal System Output
Professor Doctor Sebastian Moller
Quality and Usability Lab
Technische Universitat Berlin
Text-To-Speech Synthesis

(Refer Slide Time: 00:18)

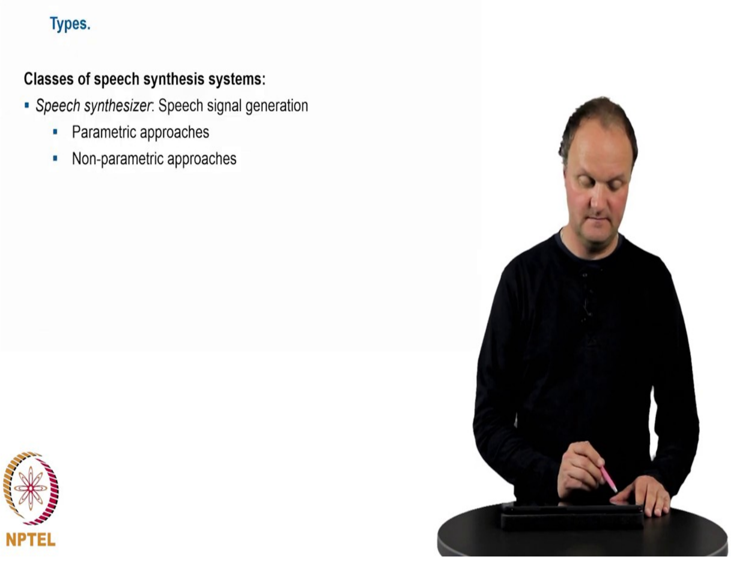
Text-To-Speech Synthesis

Prof. Dr.-Ing. Sebastian Möller
Quality and Usability Lab
Technische Universität Berlin

RWTH AACHEN UNIVERSITY
TU9
MOOC@TU9
www.tu9.de/mooc
TECHNISCHE UNIVERSITÄT DRESDEN
TUM
Technische Universität München

This chapter of the course will deal with another application of speech acoustics namely the synthesis of speech from text. We call this text to speech synthesis or TTS. Actually it seems to be quite the opposite process of speech recognition but the problems which we have are slightly different from the ones we have to deal with in speech recognition.

(Refer Slide Time: 00:43)



Types.

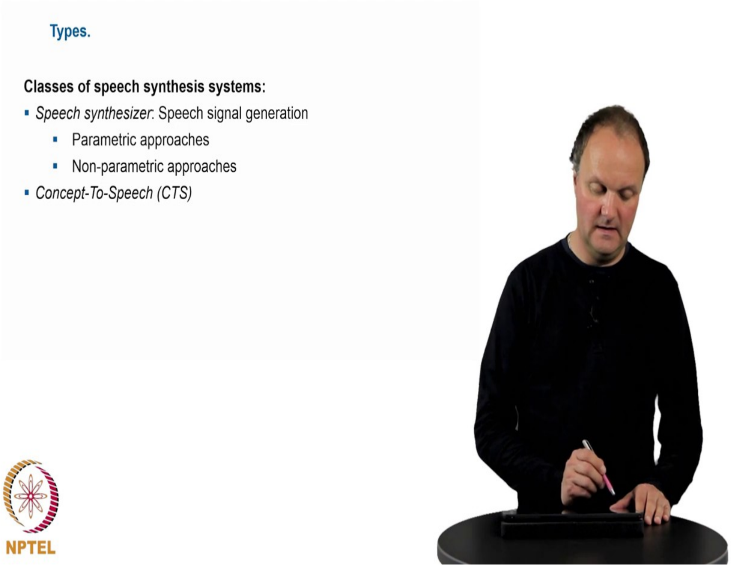
Classes of speech synthesis systems:

- *Speech synthesizer*: Speech signal generation
    - Parametric approaches
    - Non-parametric approaches

The different types of speech synthesizers which are generally called synthesizers but which do not necessarily perform the same function. The function which has to be dealt with by all types of synthesizers is the speech signal generation. Actually we want to have a speech signal produced from a text by a computer.

And there are different parametric and non-parametric approaches to deal with this problem which I will explain later on in this course.

(Refer Slide Time: 01:16)



Types.

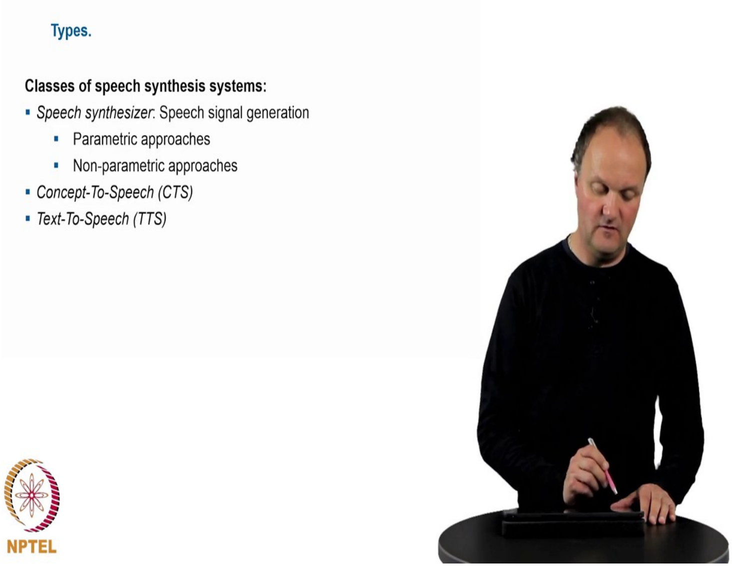
Classes of speech synthesis systems:

- *Speech synthesizer*: Speech signal generation
  - Parametric approaches
  - Non-parametric approaches
- *Concept-To-Speech (CTS)*

Now the differences arise when we take the input information for the synthesis process. So there are some synthesizers which start from a concept of the speech to be produced, which may include also non-autographic information for example where to put intonations, which templates to use and so on.

This is a more easy process than to

(Refer Slide Time: 01:43)



Types.

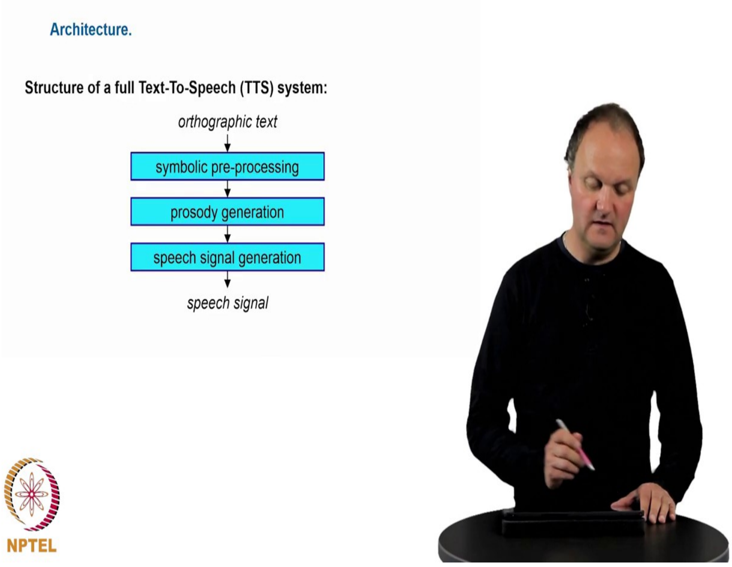
Classes of speech synthesis systems:

- *Speech synthesizer*: Speech signal generation
  - Parametric approaches
  - Non-parametric approaches
- *Concept-To-Speech (CTS)*
- *Text-To-Speech (TTS)*

extract all that information from written text. This is the most general task to produce the speech signal out of purely orthographic text. This is what we call the text-to-speech synthesis and this will be the part of synthesis we will deal with in the following.

In order to do

(Refer Slide Time: 02:06)

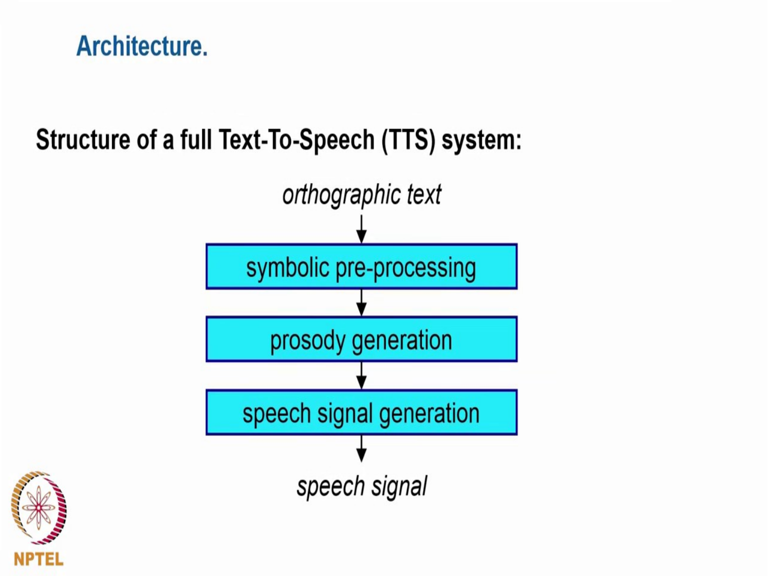


Architecture.

Structure of a full Text-To-Speech (TTS) system:

synthesis from text we have to deal with 3 separate problems. The first is called the symbolic preprocessing,

(Refer Slide Time: 02:14)



Architecture.

Structure of a full Text-To-Speech (TTS) system:

orthographic text

symbolic pre-processing

prosody generation

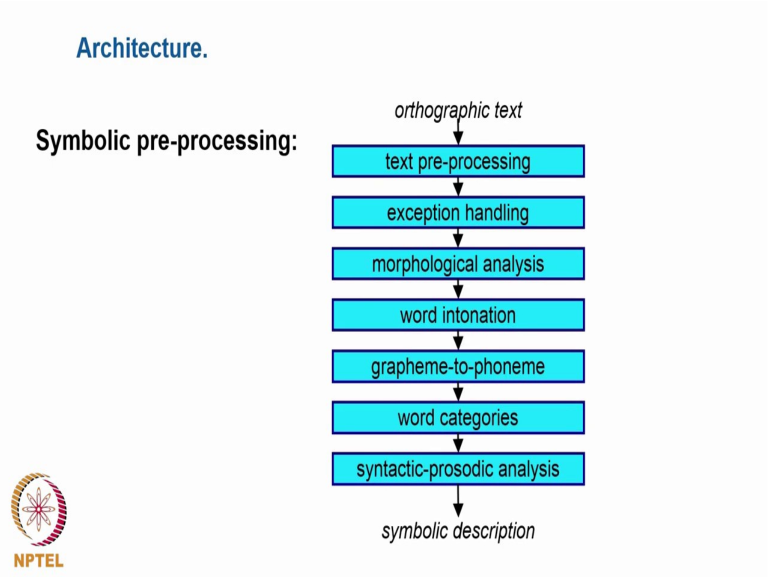
speech signal generation

speech signal

the second the prosody generation and the third the speech signal generation.

Let us start with the symbolic preprocessing.

(Refer Slide Time: 02:24)



Architecture.

Symbolic pre-processing:

orthographic text

text pre-processing

exception handling

morphological analysis

word intonation

grapheme-to-phoneme

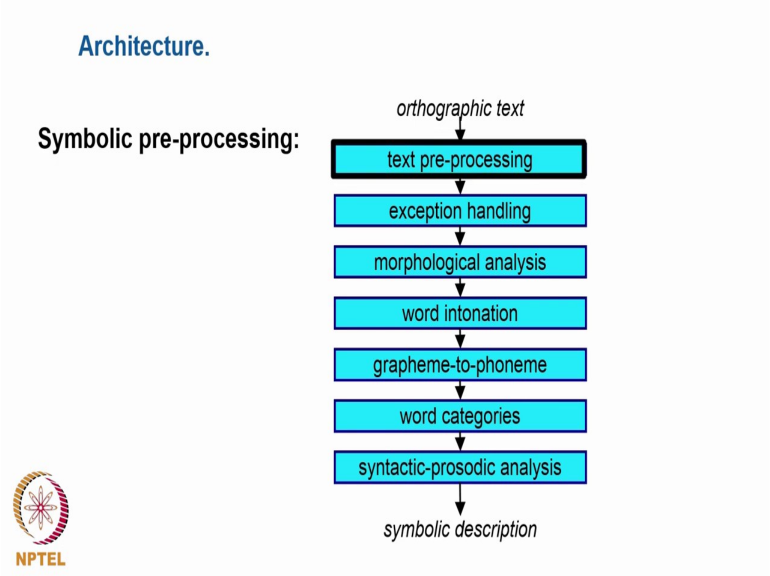
word categories

syntactic-prosodic analysis

symbolic description

And let us assume that we want to read, for example the text of an email.

(Refer Slide Time: 02:30)



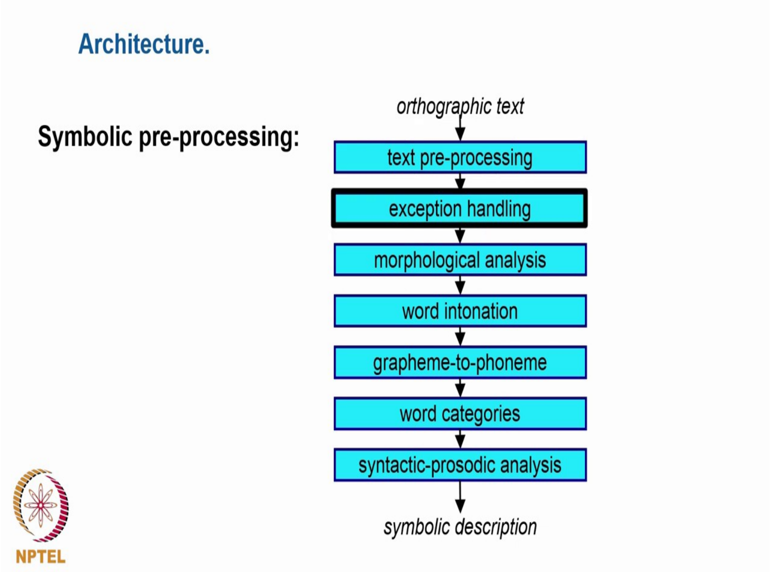
Architecture.

Symbolic pre-processing:

This text usually contains number of symbols which are not typical text, for example the time when the email has been sent, then some abbreviations which have to be resolved and so on.

All this non-orthographic text has to be translated into orthographic text in order for the speech synthesizer to pronounce it.

(Refer Slide Time: 02:56)

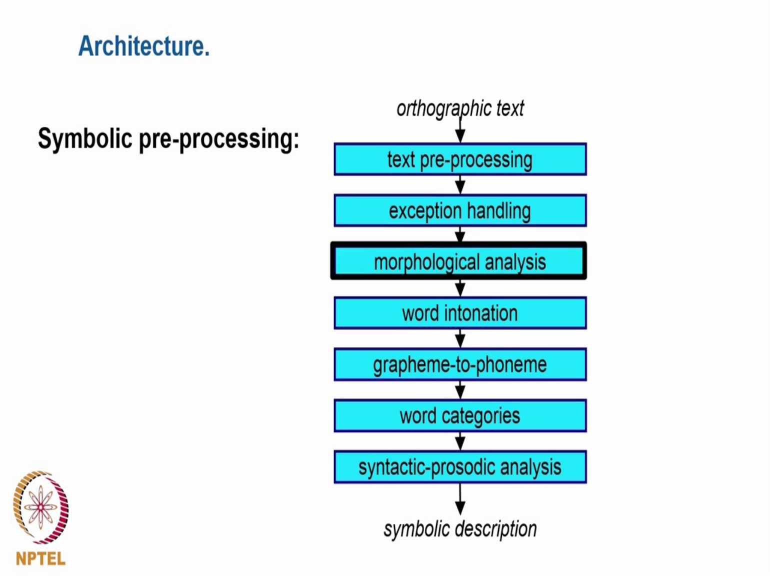


Architecture.

Symbolic pre-processing:

We might also have to deal with some exceptions, for example foreign words which have to be pronounced in a different way, according to different rules than the normal standard language.

(Refer Slide Time: 03:09)

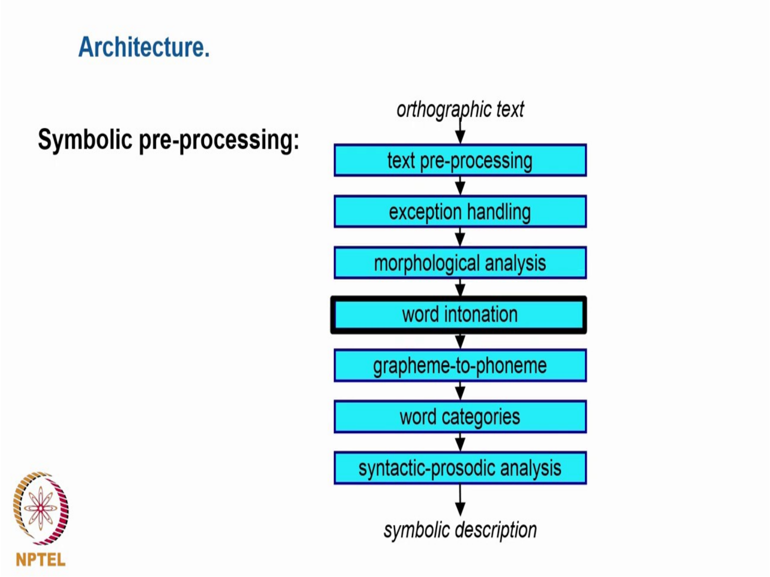


Architecture.

Symbolic pre-processing:

Then the words which are written down have to be analyzed with respect to their morphology. For example finding the core of the words, the suffixes, the endings, the prefixes which may also mean something for the word, decomposition into those constituents is helpful in order to determine pronunciation of that particular word.

(Refer Slide Time: 03:39)

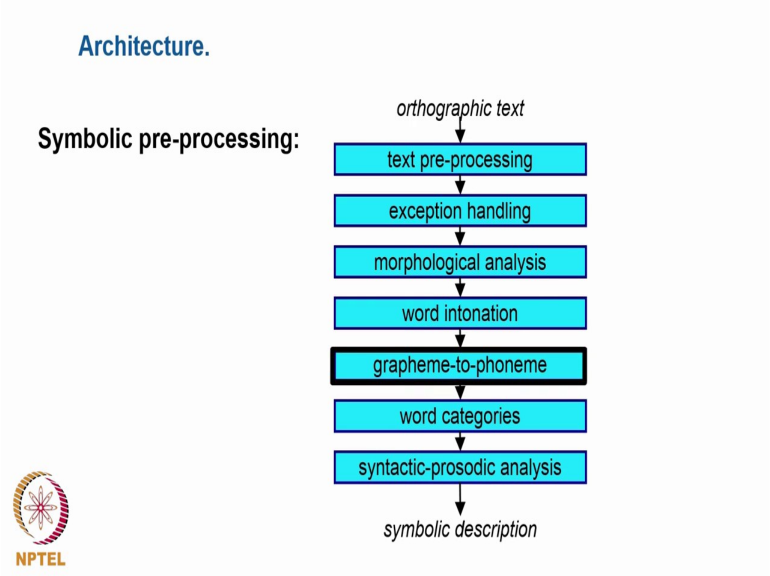


Architecture.

Symbolic pre-processing:

This decomposition also helps in the finding of the right intonation that it is the accents of the words so where the word is stressed.

(Refer Slide Time: 03:53)



Architecture.

Symbolic pre-processing:

orthographic text

text pre-processing

exception handling

morphological analysis

word intonation

grapheme-to-phoneme

word categories

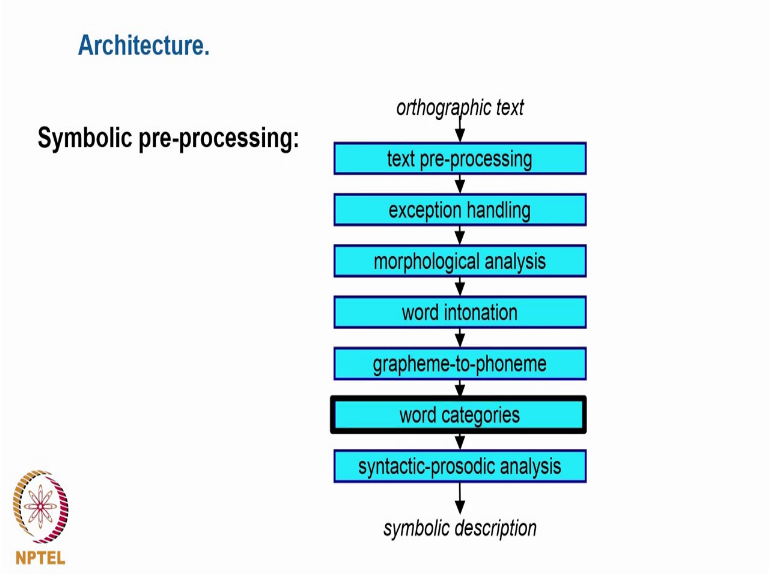
syntactic-prosodic analysis

symbolic description

NPTEL

And then the individual constituents of the word can be translated from the orthographic transcription, from the so-called graphemes to the sound description, to the so-called phonemes which constitute the pronunciation of the words.

(Refer Slide Time: 04:12)



Architecture.

Symbolic pre-processing:

orthographic text

text pre-processing

exception handling

morphological analysis

word intonation

grapheme-to-phoneme

word categories

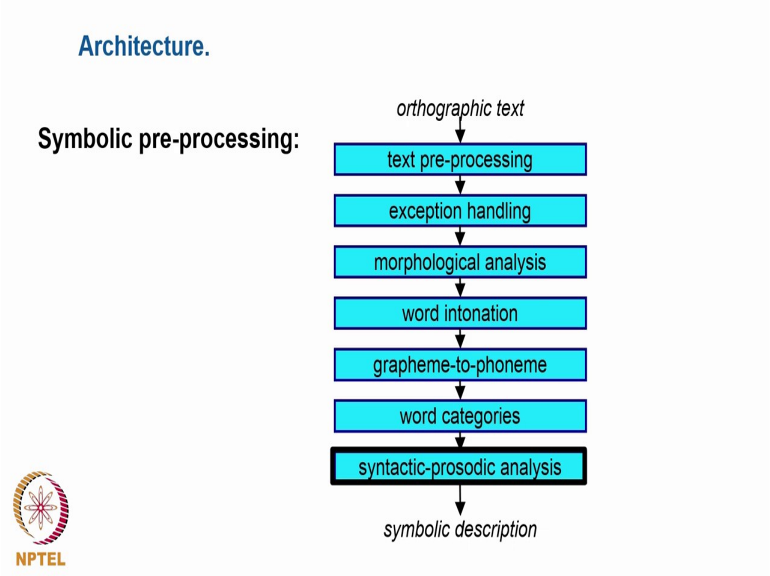
syntactic-prosodic analysis

symbolic description

NPTEL

In order to deal with longer units it is sometimes meaningful to extract the categories of certain words. So for example, what is the noun, what is the verb, or what is an object, in order to find the right pronunciation of the entire sentence.

(Refer Slide Time: 04:32)



Architecture.

Symbolic pre-processing:

orthographic text

text pre-processing

exception handling

morphological analysis

word intonation

grapheme-to-phoneme

word categories

syntactic-prosodic analysis

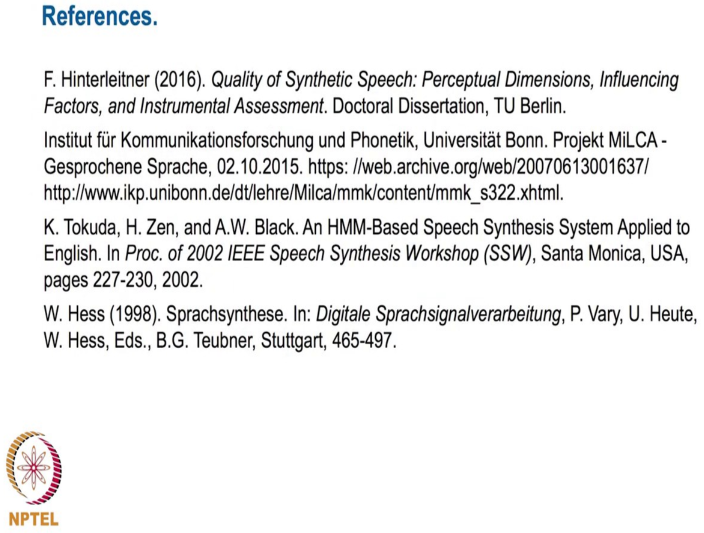
symbolic description

Ideally we would have a semantic meaning analysis of a full sentence but in most cases this would not be possible. This is why the analysis further is restricted to the syntactic characteristics so what makes the rules for producing sentences and how these are translated into prosodic cues.

The prosody, as you have learnt in the beginning, includes 3 physical parameters, the amplitude, the fundamental frequency and the length of certain sounds. And all those have to be produced in the correct way so that the produced language sounds natural in the end.

The information which has been collected in all these individual modules forms the basis for the second step which is the actual production rules for the prosody and then for the third and final step, that is the generation of the actual speech signal.

(Refer Slide Time: 05:39)



References.

F. Hinterleitner (2016). *Quality of Synthetic Speech: Perceptual Dimensions, Influencing Factors, and Instrumental Assessment*. Doctoral Dissertation, TU Berlin.

Institut für Kommunikationsforschung und Phonetik, Universität Bonn. Projekt MiLCA - Gesprochene Sprache, 02.10.2015. https: //web.archive.org/web/20070613001637/ http://www.ikp.unibonn.de/dt/lehre/Milca/mmk/content/mmk_s322.xhtml.

K. Tokuda, H. Zen, and A.W. Black. An HMM-Based Speech Synthesis System Applied to English. In *Proc. of 2002 IEEE Speech Synthesis Workshop (SSW)*, Santa Monica, USA, pages 227-230, 2002.

W. Hess (1998). Sprachsynthese. In: *Digitale Sprachsignalverarbeitung*, P. Vary, U. Heute, W. Hess, Eds., B.G. Teubner, Stuttgart, 465-497.