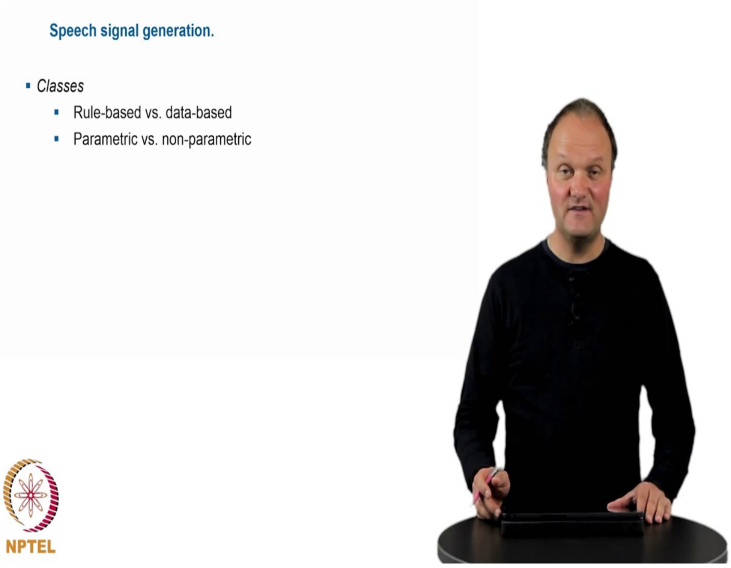
Multimodal System Output
Professor Doctor Sebastian Moller
Quality and Usability Lab
Technische Universitat Berlin
Speech Generation

(Refer Slide Time: 00:15)

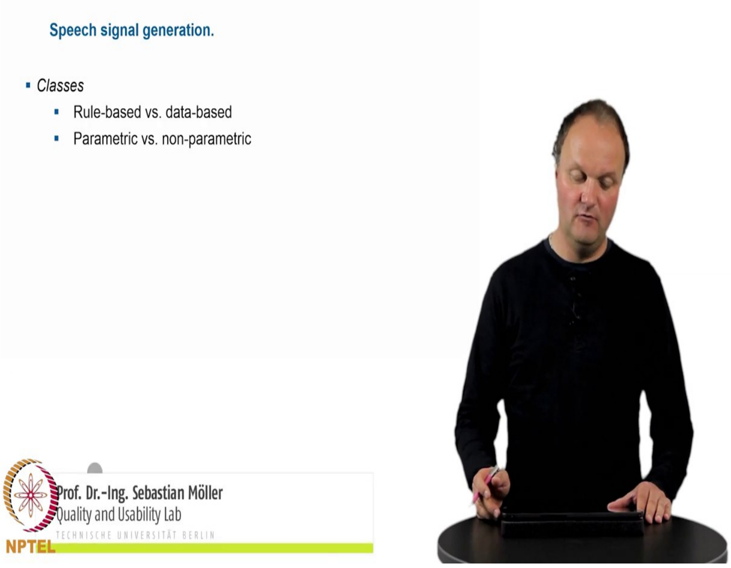


Speech signal generation.

- *Classes*
  - Rule-based vs. data-based
  - Parametric vs. non-parametric

As mentioned before there are different principles for the actual generation of the speech signal. They may be

(Refer Slide Time: 00:21)



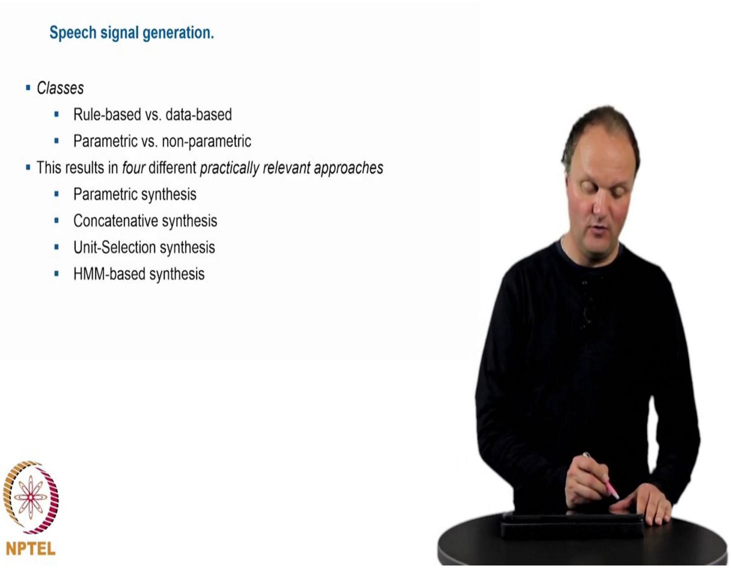
Speech signal generation.

- *Classes*
  - Rule-based vs. data-based
  - Parametric vs. non-parametric

Prof. Dr.-Ing. Sebastian Möller
Quality and Usability Lab
TECHNISCHE UNIVERSITÄT BERLIN
NPTEL

rule-based or data-based; they may be parametric or non-parametric. But in practice there are actually 4

(Refer Slide Time: 00:29)



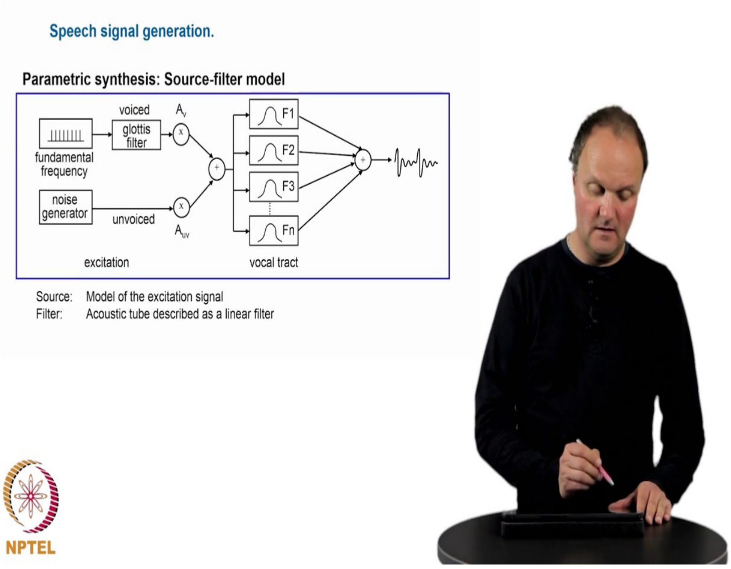
Speech signal generation.

- *Classes*
  - Rule-based vs. data-based
  - Parametric vs. non-parametric
- This results in *four* different *practically relevant approaches*
  - Parametric synthesis
  - Concatenative synthesis
  - Unit-Selection synthesis
  - HMM-based synthesis

principles which have proven to be practically relevant and to produce naturally sounding speech signals.

These are the parametric synthesis, the concatenative synthesis which comes in two flavors, first called the unit selection synthesis and finally the synthesis approach which is based on Hidden Markov Models HMMs.

(Refer Slide Time: 00:53)



Speech signal generation.

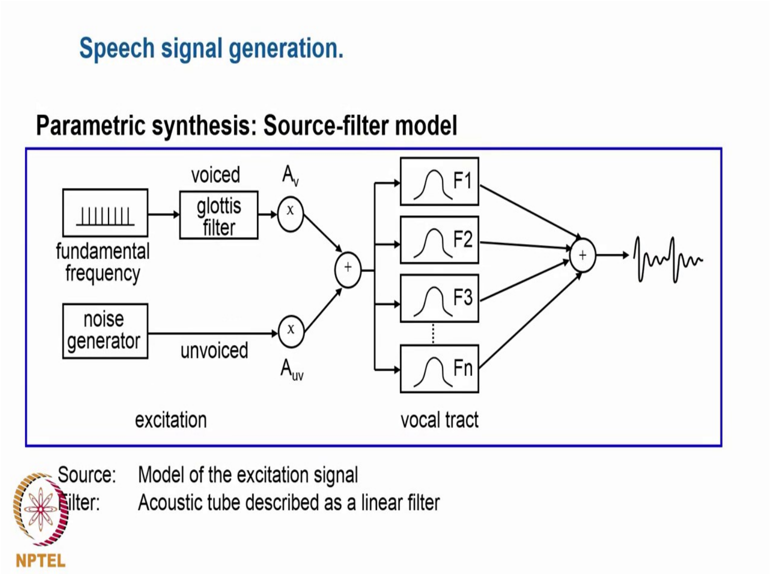
Parametric synthesis: Source-filter model

Source: Model of the excitation signal
Filter: Acoustic tube described as a linear filter

The first such approach, the parametric synthesis is the one which you already know from chapter 6 of this course because it is mainly based on the source-filter model of speech production.

The idea is to identify all

(Refer Slide Time: 01:07)



Speech signal generation.

Parametric synthesis: Source-filter model

Source: Model of the excitation signal

Filter: Acoustic tube described as a linear filter

NPTEL

the parameters which are used in the source-filter model for producing a particular sound and putting values which correspond to that particular sound.

So we first need to differentiate whether this is a periodic sound, then we need to decide about fundamental frequency that is including the glottis filter, or if it is a noisy sound, then the noise signal to be generated.

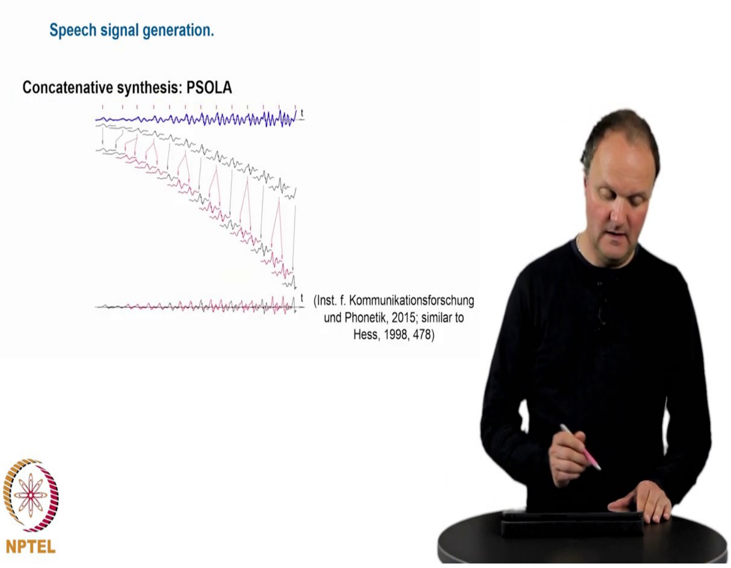
We can amplify those excitation signals and add them in case we have a mixed excitation and then transmit everything through the vocal tract filter which here is modeled as a parallel structure of different formant filters.

As you know these formants are important for the characteristics of the sound to be produced. So for each sound to be produced we have to put values for all these characteristics of the speech production process.

Unfortunately, this source filter model is a rather simplified version of the speech production process and this is why the speech signals which come out of such a process are relatively limited with respect to their naturalness.

In order to deal with that problem,

(Refer Slide Time: 02:22)



Speech signal generation.

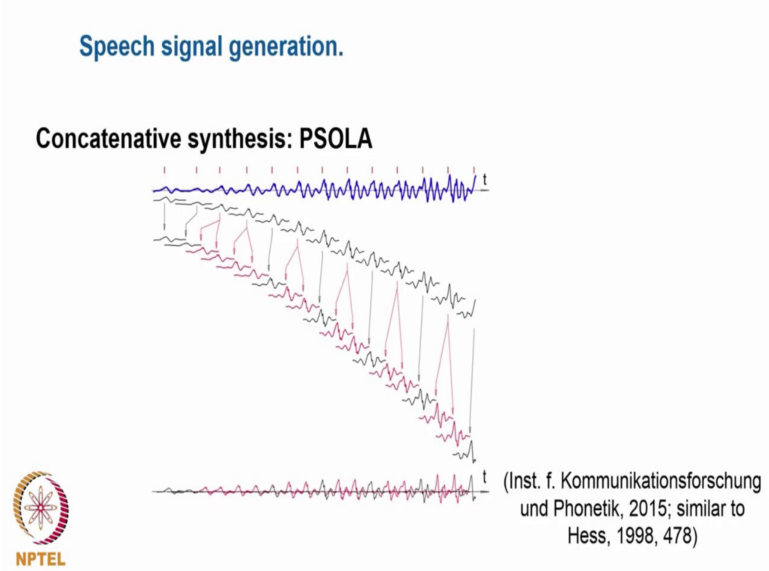
Concatenative synthesis: PSOLA

(Inst. f. Kommunikationsforschung und Phonetik, 2015; similar to Hess, 1998, 478)

people have thought of recording speech signals from real speakers, from natural speakers and concatenating them in a way which allows still the manipulation of different prosodic characteristics like the amplitude, the fundamental frequency and the length of the individual sounds but still using natural speech which should sound more natural than the one produced with the source-filter model.

We usually take units

(Refer Slide Time: 02:56)



Speech signal generation.

Concatenative synthesis: PSOLA

(Inst. f. Kommunikationsforschung und Phonetik, 2015; similar to Hess, 1998, 478)

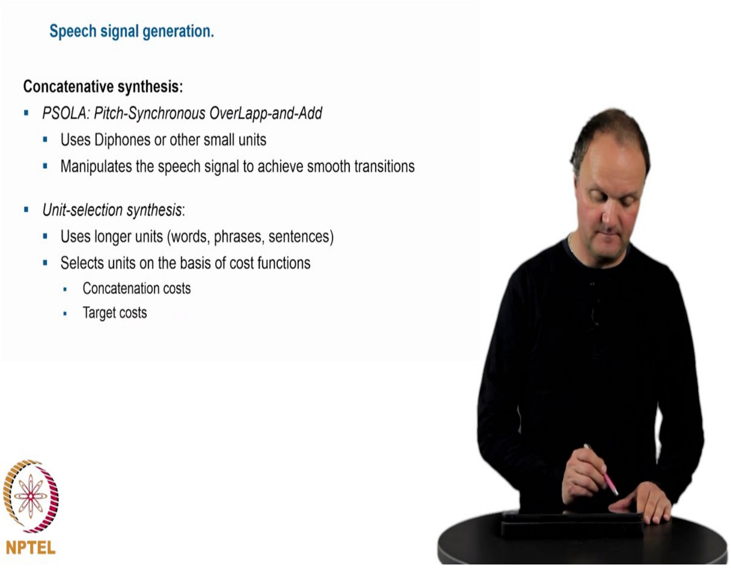
which correspond to transitions of different phons, as you see here in this example and these units are cut out in a pattern which is equivalent to the pitch that is the fundamental frequency when the glottis is open is marked here first and then a little segment of this signal is cut out corresponding to this pitch period.

These individual segments can then be manipulated with respect to their amplitude, for example in order to produce the intonation, with respect to their fundamental frequency, for example by putting 2 segments instead of 1 and of course also by, with respect to their length, by just using more of those segments.

And they are then overlapped and add to form a new speech signal which has different characteristics, for example here you see that the fundamental frequency or the period corresponding to the fundamental frequency is smaller, the fundamental frequency is higher than the original sample here.

We call this process the pitch synchronous overlap and add or PSOLA algorithm for concatenating speech units.

(Refer Slide Time: 04:21)



Speech signal generation.

Concatenative synthesis:

- *PSOLA: Pitch-Synchronous OverLapp-and-Add*
  - Uses Diphones or other small units
  - Manipulates the speech signal to achieve smooth transitions
- *Unit-selection synthesis*:
  - Uses longer units (words, phrases, sentences)
  - Selects units on the basis of cost functions
    - Concatenation costs
    - Target costs

And it is usually based on transitions, and on small units like diphones and it involves a lot of manipulation of the speech signal which then unfortunately results also in quite some unnaturalness of the resulting speech signal.

People have tried to avoid this unnaturalness by putting less manipulation into the speech signals and this can be achieved by taking longer units for example words or phrases or sentences.

If we concatenate those longer units we have to find the ones which are fitting best to the speech signal to be produced but also which fits best among themselves, which means that this synthesis approach is based on the calculation of the so-called cost function which consists of concatenation costs, costs that the units fit amongst themselves and the so-called target costs that is the costs that the unit which is used for the synthesis actually fits what needs to be synthesized.

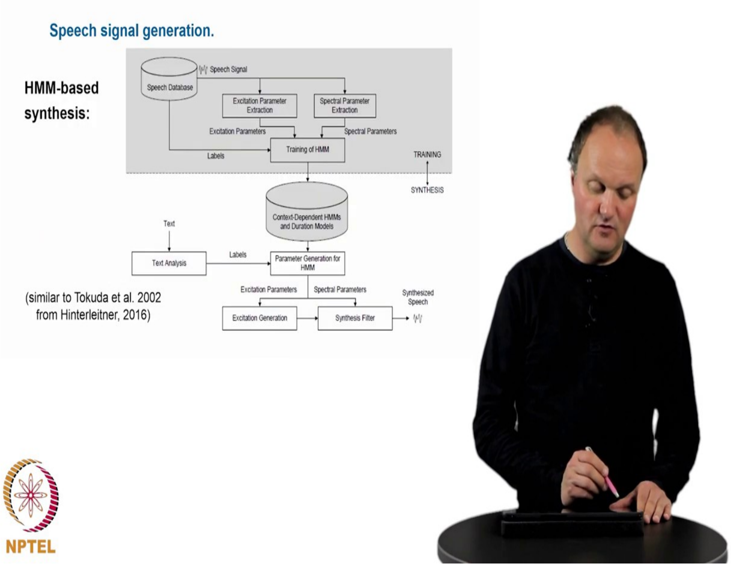
So the principle of the so-called unit selection synthesis is to find an ideal sequence of units which minimizes the costs in terms of concatenation costs and target costs.

The quality which can be achieved by this unit selection principle can be relatively high in case that the units which are in the inventory of this synthesizer fit well to the units which need to be used in order to produce the text which is desired, that is the quality depends quite a lot on which text you would like to synthesize and it also depends quite a lot on how many and which units you have in your inventory.

This makes usually the inventory of such a synthesizer a relatively large one. So this is a synthesis process with a rather large footprint requiring lots of memory.

In case that you do not

(Refer Slide Time: 06:35)



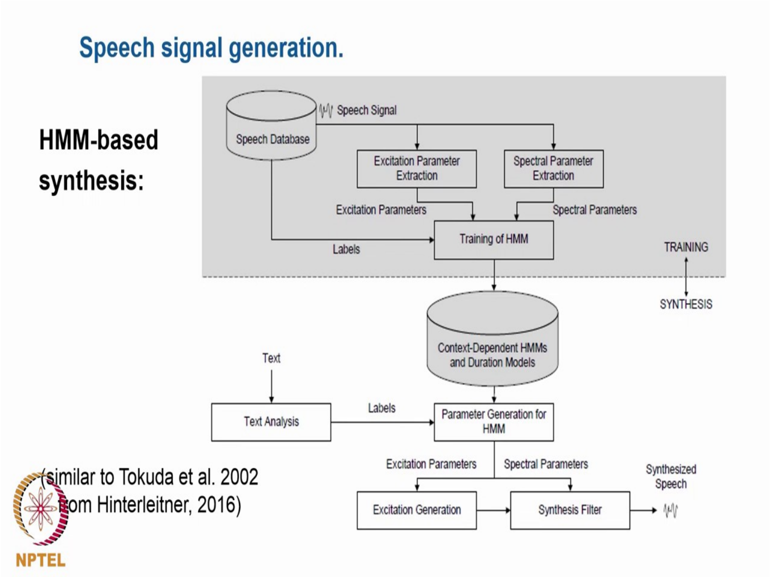
Speech signal generation.

HMM-based synthesis:

(similar to Tokuda et al. 2002 from Hinterleitner, 2016)

have such a large memory available you can make use

(Refer Slide Time: 06:39)



Speech signal generation.

HMM-based synthesis:

(similar to Tokuda et al. 2002 from Hinterleitner, 2016)

of the fourth principle which is based on the Hidden Markov Models.

As we, you know from automatic speech recognition, Hidden Markov Models can be used to find a path between states; and this path that can be optimized by calculating the probabilities.

Something very similar happens if you want to concatenate the speech units, you can try to find the path across a parameter space and then select the parameters which optimally fit your synthesis task. This is done in 2 steps.

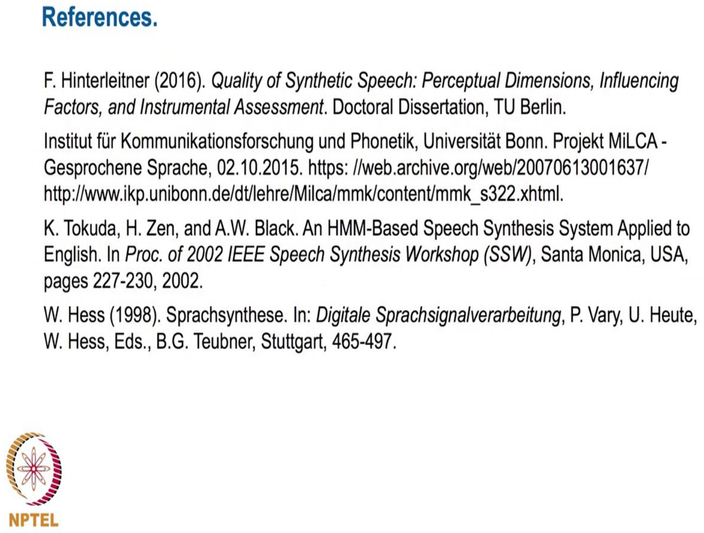
You first have to train the Hidden Markov Model in terms of providing a speech database and training parameters related to the excitation process and to the vocal shaping process which correspond to individual sounds which are included in the labels here. And then put that information into a trained Hidden Markov Model.

In the second step during the synthesis, the text is analyzed, labels are extracted and then this Hidden Markov Model helps you to generate an optimum set of parameters both for the excitation and for the vocal shaping. And then these parameters can be used in a standard parametric speech synthesis process.

Results from this Hidden Markov Model based synthesis process are rather good, although they might not necessarily reach the quality of a unit selection synthesizer, depending of course on the inventory of that unit selection synthesizer.

The big advantage of the Hidden Markov based approach is that the inventory is much smaller. The footprint of this synthesizer is much smaller than the one of unit selection synthesis.

(Refer Slide Time: 08:45)



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.