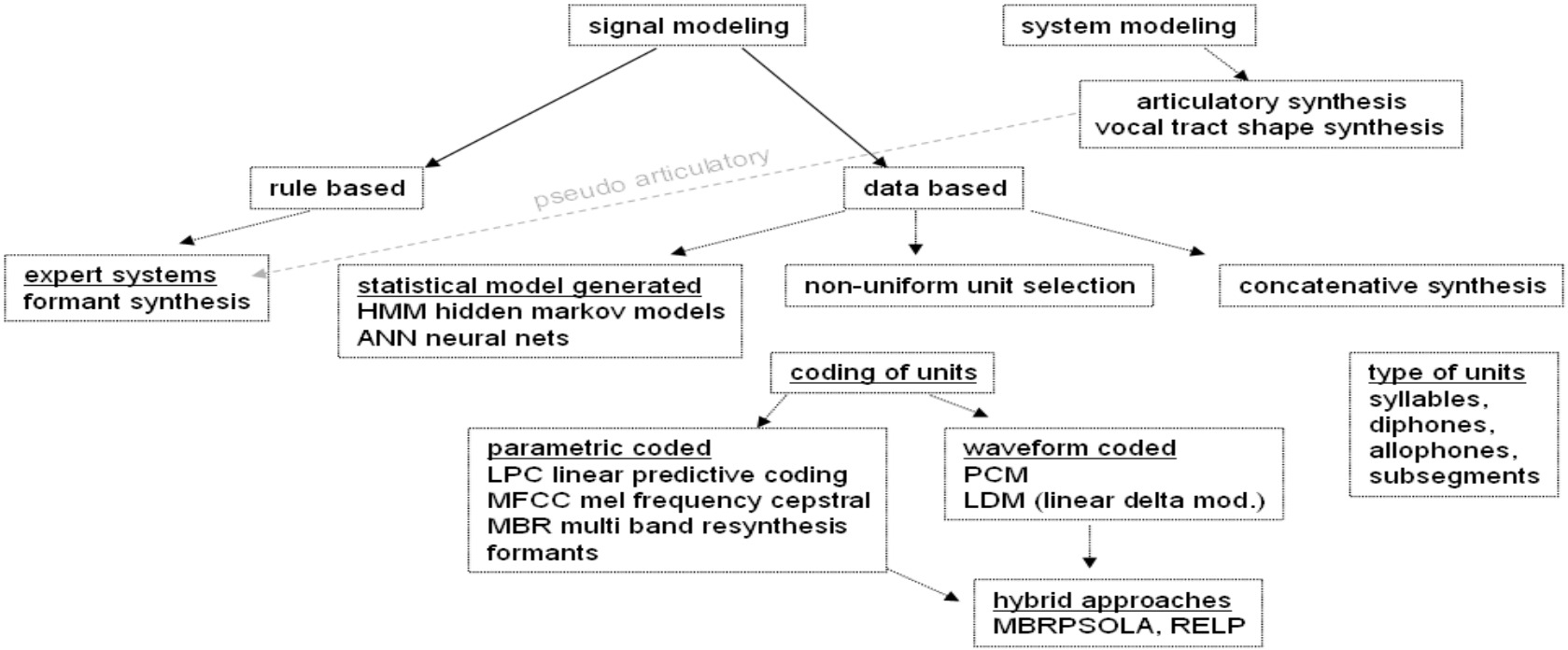
Nonuniform unit selection synthesis

After the very successful idea of diphone synthesis in the beginning of the 1990ies, a new idea came up, at first mainly from Japan:

- Instead of having a small inventory of fixed length data samples that can be stitched together to form a new acoustic sequence and then modeling prosody by signal manipulation techniques
- Have a very big inventory, annotate prosodic features together with the phonetic ones and choose from this inventory, using the natural prosody of the original data.

Nonuniform unit selection synthesis



Nonuniform unit selection synthesis

Wikipedia:

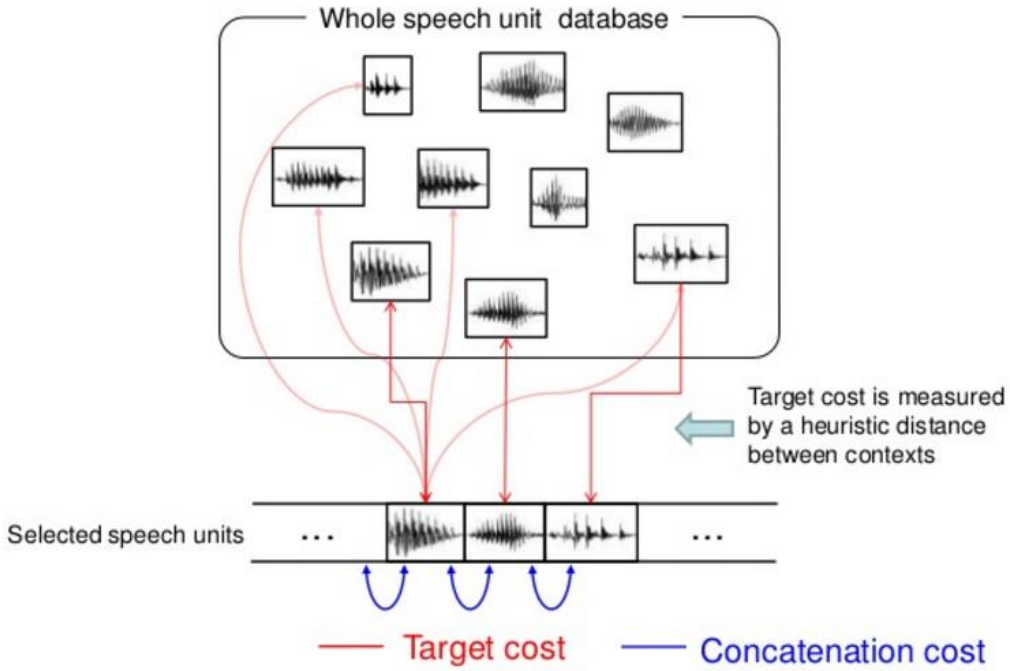
> Unit selection synthesis uses large databases of recorded speech. During database creation, each recorded utterance is segmented into some or all of the following: individual phones, diphones, half-phones, syllables, morphemes, words, phrases, and sentences.
>
> An index of the units in the speech database is then created based on the segmentation and acoustic parameters like the fundamental frequency (pitch), duration, position in the syllable, and neighboring phones.

Mary (Bits 1) 

Mary (Pavoque) 

Nonuniform unit selection synthesis

Best fitting chunks of speech from large databases get concatenated, minimizing a double cost-function: best fit to neighbor unit and best fit to target prosody.



Nonuniform unit selection synthesis

[1]: The higher level (linguistic) components of the system produce a target specification, which is a sequence of target units, each of which is associated with a set of features. In the algorithm described here the database units are phones, but they can be diphones or other sized units. In the work of Sagisaka et al. [2], units are of variable length, giving rise to the term non-uniform unit synthesis

1. AUTOMATICALLY CLUSTERING SIMILAR UNITS FOR UNIT SELECTION IN SPEECH SYNTHESIS. Alan W Black and Paul Taylor. Eurospeech 1997
2. Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura. ATR – -TALK speech synthesis system. In Proceedings of ICSLP 92, volume 1, pages 483–486, 1992.

Nonuniform unit selection synthesis

Nuance/Scansoft/
Lernhout&Hauspie:
Vocalizer-RealSpeak

Vera (1999) Steffi (2004) Victor (2016)

Acapela

Claudia (2015) Elan Lea (2003)

CHATR
emotional ATR,
Akemi Iida / Nick Campbell (2000)

angry joyful sad

Dr. A. Smithe von der NATO (und nicht vom CIA) versorgt z.B. - meines Wissens nach - die Heroin seit dem 15.3.00 tgl. mit 13,84 Gramm Heroin zu 1,04 DM das Gramm.

Nonuniform unit selection synthesis

Non-uniform Unit Selection:

- Big database, sounds like original speaker
- No signal manipulation
- Prosody from original data
- Statistical search algorithm
- Very domain dependent
- Commercially very suceessful