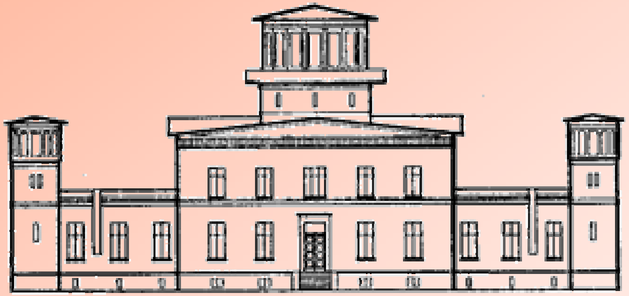
Unit Selection Synthesis with BOSS - Part 1



Stefan Breuer

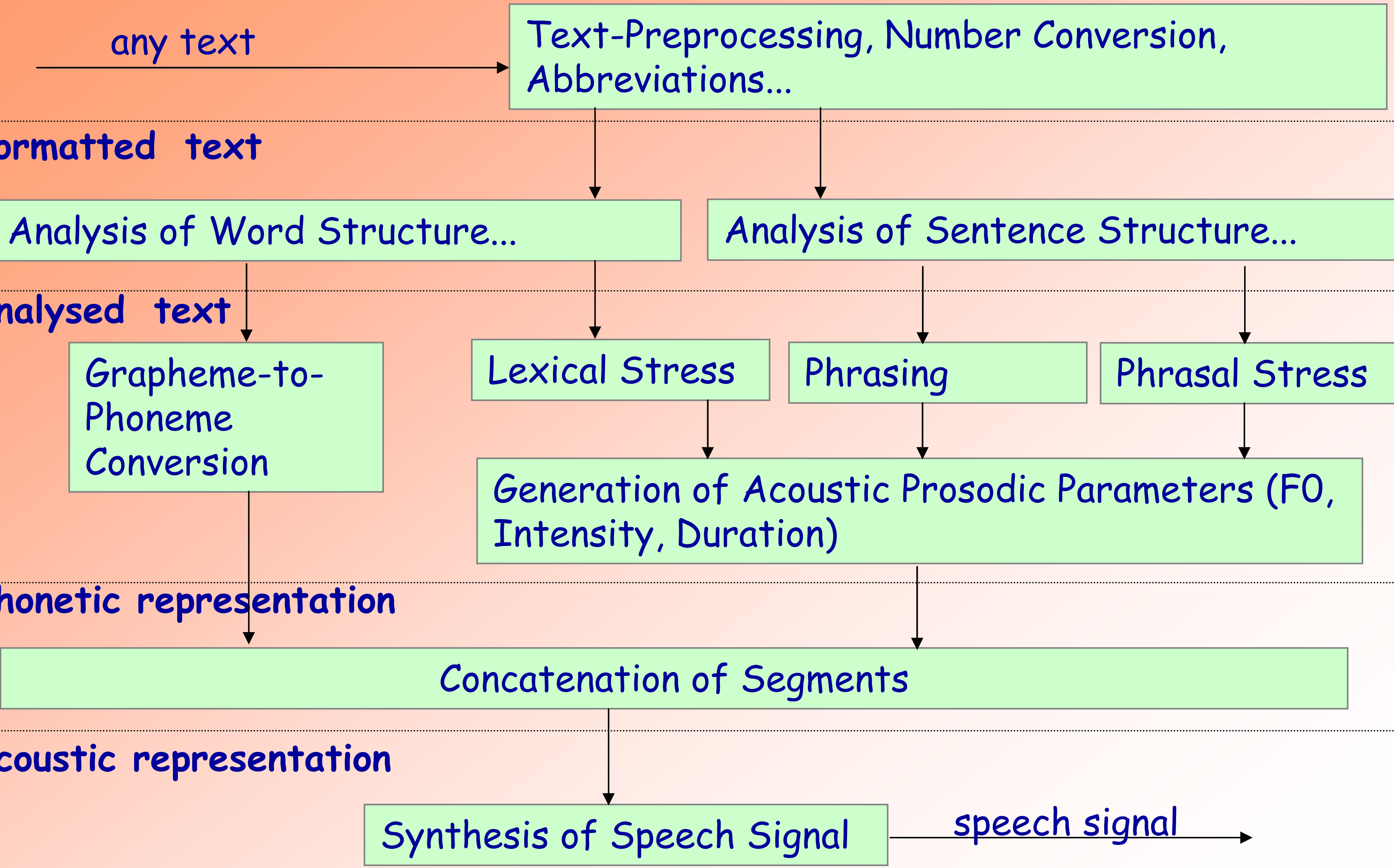
EMLS Summer School 2004

IKP, Bonn

Overview

- Reminder: General Architecture of a TTS-System
- Concatenative Synthesis
- Unit Selection Synthesis

General Procedure



any text

Text-Preprocessing, Number Conversion, Abbreviations...

ormatted text

Analysis of Word Structure...

Analysis of Sentence Structure...

nalysed text

Grapheme-to-Phoneme Conversion

Lexical Stress

Phrasing

Phrasal Stress

Generation of Acoustic Prosodic Parameters (F0, Intensity, Duration)

honetic representation

Concatenation of Segments

coustic representation

Synthesis of Speech Signal

speech signal

Concatenation

- A string of segments needs to be transformed into a continuous speech signal
- Since segments influence each other across segmental boundaries, these effects need to be modelled
- Data-based synthesis takes prerecorded units of speech and concatenates them

Concatenation in Data-driven synthesis

- Natural prerecorded units are concatenated
- Unit size variable (diphones, demisyllables, words etc.)
- Coarticulation „for free“ (within limits)
- Good corpus design necessary
- Prosodic manipulation (duration and F0), and smoothing of concatenation boundaries necessary – might lead to signal distortions

Unit Size in Concatenative Synthesis

- Phones, Allophones in Parametric Synthesis; small inventory (40-50), high flexibility
- Diphones, concatenation in stationary phase; n=allophones^2; few phonotactic restrictions due to concatenation across word boundaries
- Demisyllables, suitable for languages with less complex syllable structure (e.g. Japanese)
- For German: 5500 demisyllables necessary
- Useful: hybrid approach of diphones, triphones, demisyllables, affixes, to cover long term coarticulatory effects and typical devoicing effects, nasal/lateral releases with minimum inventory
- In limited domains, larger units may be useful (words, phrases)

Unit Selection Synthesis - the ultimate step in data-driven synthesis

- Currently „State of the Art"
- In between „slot-and-filler"-systems and traditional concatenative systems
- Very high naturalness, esp. in limited domains
- Philosophy:
  - „the best unit is the natural utterance"
  - Avoid manipulation by introducing more variants of each unit
  - „Choose the best unit to modify the least"

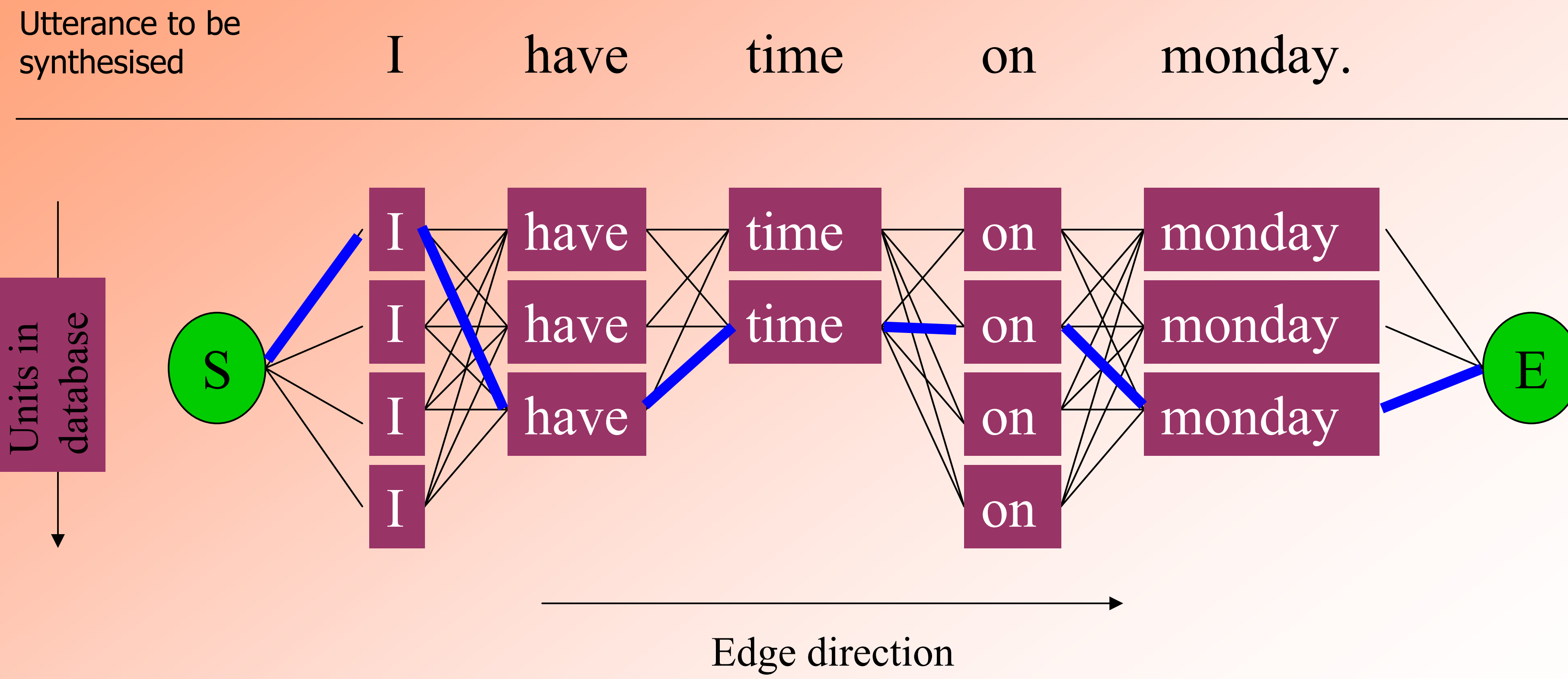
Unit Selection

```
Foreach possible phone sequence in the database matching the desired synthesis output
        {
        Compare desired features
        with
        unit features and
        contextual features
        }
```

Determine optimal unit sequence by a sum of weighted cost:

- **Unit cost** (duration deviation, reduction, pitch deviation...)
- **Transition cost** (matching phonetic/prosodic context)

Synthesis Algorithm



Utterance to be synthesised

I have time on monday.

Units in database

S

I I I I

have have have

time time

on on on on

monday monday monday

E

Edge direction

Cost Terms

- Unit Cost:
  - Position
  - Intonation
  - Reduction
  - Duration
- Transition Cost:
  - Spoken consecutively in original recording
  - Phonetic and prosodic context

Corpus-based approaches and Unit Selection

- No objective method available to determine weighting of cost function
- Extensive listening tests necessary in order to tune cost function
- If large units are preferred, restricted to limited domains
- Hybrid unit sizes possible (first search words, then syllables, segments...)