Introduction to Speech Synthesis

Petra Wagner

IKP – University of Bonn, Germany

Vienna - ESSLLI 2003

The goal ...

- Transformation of written text or semantic/pragmatic concepts into speech in a natural way
- ...but
  - What is adequate?
  - What is natural? (imitation of human speaker?)

What is the problem?

„***Apologies for multiple postings***

Dear ISCA members,

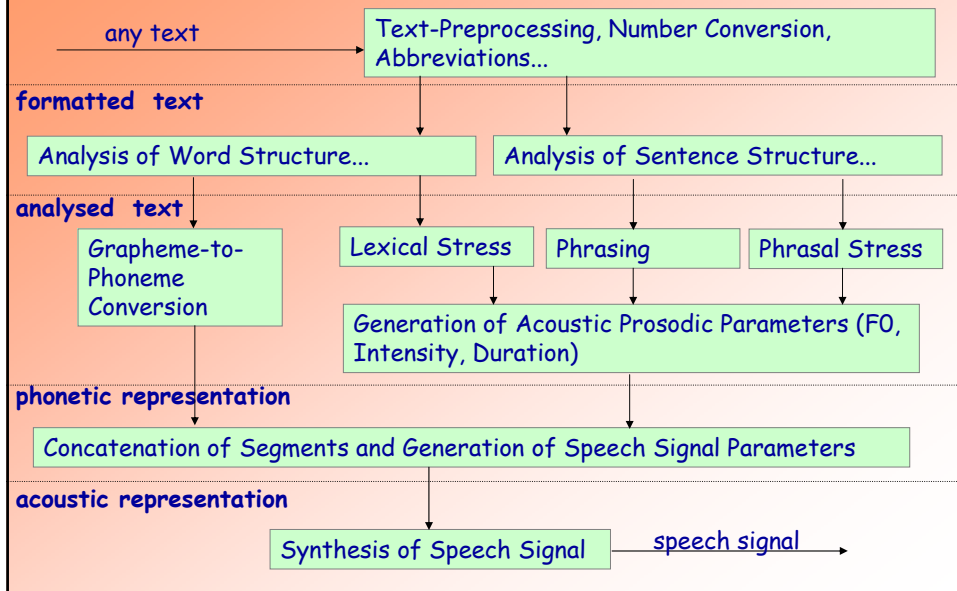The 22nd West Coast Conference on Formal Linguistics (WCCFL XXII) will be held on March 21-23, 2003, at the University of California, San Diego. Abstracts from all areas of formal linguistics are invited for 20-minute talks in the general session."

„Ciao Petra. Thanks for the invitation :-) My train will arrive approx. 8:15 in Cologne-could you tell me asap whether you will be able to pick me up at the station? Arrivederci."

Overview

- General Architecture of a TTS-System
- Symbolic Preprocessing
- From Segments to SynthesisUnits or Acoustic Parameters
- Acoustic Synthesis
- Next Generation: Corpus Based Synthesis
- Evaluation of Synthesis Systems

General Procedure

any text → Text-Preprocessing, Number Conversion, Abbreviations...

formatted text

| Analysis of Word Structure... | Analysis of Sentence Structure... |

analysed text

Grapheme-to-Phoneme Conversion

| Lexical Stress | Phrasing | Phrasal Stress |

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

phonetic representation

Concatenation of Segments and Generation of Speech Signal Parameters

acoustic representation

Synthesis of Speech Signal → speech signal

Problems to solve in text preprocessing...

- Cardinal, ordinal, other numbers

- Pronunciation of abbreviations

- Ambiguous punctuation marks, emoticons, diacritics etc.

Problems in grapheme-2-phoneme conversion and lexical stress assignment

• Loan words

• Proper Names

• Inflectional/compositional morphology (morphophonology)...

phones + lexical stress

Problems to solve in sentence analysis...

• Aim: Determination of the stressed words+prosodic phrase boundaries and type (falling vs. rising, finality)
• Syntactic analysis difficult, esp. with fragmentary or complex input
• Influence of context on prosody
• No 1:1-mapping of syntax:prosody
• Usually: POS+punctuation marks

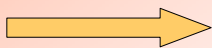Prosodically annotated text

Prosodic Annotation

```
DIs(12)              And
Iz(16)               Dis H*
@n(2)                @
Ig(8)                n'V H*-L
z'a:m(20)            D@
pl=(0)               wVn L*-L%
.
```

→ Prosodically annotated text

Solution in Concept-to-Speech?...

- Semantic Focus
- Domain-specific knowledge about topic, speaking style
- Text History, Context
- Contrast, Emphasis
- Deaccentuation of contextually given material

→ CTS only in limited domains

Generation of Acoustic Prosodic Parameters

- Important for naturalness (but also intelligibility)
- Parameters: duration, intensity, fundamental frequency
- Connected to ALL linguistic, paralinguistic, extralinguistic levels (intensity of less importance)
- A perfect prosodic representation needs semantic, pragmatic, emotive, analysis + specification of speaker characteristics

Approaches to Duration Generation

- rule-based duration generation (Klatt, 1979)
- sum-of-products model (van Santen, 1994)
- syllable-duration based generation (Campbell and Isard, 1991)

phone duration in ms

Rule-Based Duration Generation

- Context-sensitive „phonology-style" rules
- Specific order of rule application
- Phones are assigned inherent durations which are increased or decreased according to environment

DUR=((INHDUR-MINDUR)xPRCNT)/100+MINDUR

Sum-of-Products model

- Duration is calculated for each phone based on few parameters (position in syllable, position of syll in phrase, context, accentuation)
- Each parameter makes an either additive or multiplicative contribution to phone duration (sum-of-products)
- Phone duration determined with decision tree based on statistical findings in large corpora

Syllable-duration based model

- Syllable duration basic entity for generating phone duration
- Each phone has certain „elasticity", can be compressed more or less (e.g. stops less than fricatives)
- Syllable duration dependent on several factors (number of phones, nucleus type, position of syllable in phrase, lexical category of word, lexical stress...)

F_0 Generation

- Generative grammars
- Data-driven generation (e.g. neural networks, trained decision trees, statistical rules based on regression analyses)
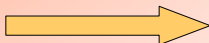
F0 values at specific reference points in signal

F_0 Generation with a generative grammar

- Often based on phonological knowledge (lexical+sentence stress, prosodic phrasing, global intonation contour) + global speed, register
- Context-sensitive rules used for predicting Fo-values of symbolic entities (2-3 points for each pitch peak or valley)
- Interpolation in between specified points
- Usually inherent declination or downstep

Data-driven F_0 Generation

• Machine-learning algorithms learn generalisations they can apply to „new" data

• necessary: prosodically annotated corpus with relevant information (durations, Fo, accentuation, boundaries...)

•Prediction of several Fo points per unit (e.g. 5 per syll)

•Nonlinear dependencies can be described with NNs

•„black box" (decision trees fairly interpretable)

F_0 Generation based on statistically built rules

- Statistical Analyses, e.g. regression analyses of a large corpus lead to isolation of factors significantly influencing Fo
- Prediction based on regression equation
- black box avoided

keep in mind...

- Every data-driven approach can only model and generalise the data you annotated (and thus believed important)
- Statistics may surprise you, but doesn't save you from studying phonetic and linguistic patterns

Concatenation-Coarticulation

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

Coarticulation
in Rule-based synthesis

- Segments 2 Acoustic Parameters (formants, antiformants, noise, quasiperiodic excitation...)
- Coarticulation Modeling based on rules based on phonetic knowledge
- Special case: Articulatory Gestures
- Maximum flexibility, phonetic production model

Acoustic parameters and duration information

Concatenation in Data-driven synthesis

- Natural prerecorded units which are concatenated
- Unit size variable (diphones, demisyllables etc.)
- Coarticulation „for free"
- Good corpus design necessary
- Prosodic manipulation, and smoothing of concatenation boundaries necessary
- If coarticulatory effects lead to a change of segmental quality, rules need to be reintroduced
- Natural sound

Units+durations+Fo-values

Segmental Units in 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
- German hybrid system: HADIFIX (HAlbsilben, DIphone, AfFIXe)

Acoustic Synthesis in Rule-based Synthesis

- Fully artificial speech signal
- Usually: formant-synthesiser
- Articulatory source-filter model
- Source signal: quasiperiodic or noise
- Linear filter models vocal tract transfer function
- Problem: all-pole filter cannot model antiformants, more complex synthesisers require more complex rule systems (Carlson 1991: 37 parameters)

Cascade/Parallel-Synthesiser by Klatt (1980)
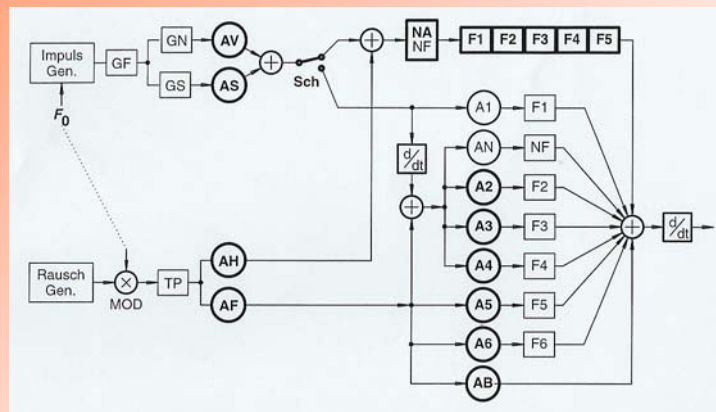
Source Signal Generation in Parametric Synthesis

- Crucial in Parametric Syntesis

- typical „buzziness"

- approaches to imitate the natural voicing appropriately (e.g. Fant's LF-model)

- female voice source difficult to model

Articulatory Synthesis: special case of parametric synthesis

- Not intended for working applications
- Prediction of articulatory configuations based on speech gestures
- Acoustic re-synthesis of gestural configurations
- Evaluation of articulatory models

Rule-based Synthesis – Pros and Cons

- Basic research (voice source parameters, articulatory phonetics, coarticulation)
- Very flexible, small allophonic inventory
- No corpus recording necessary
- Direct prosody control
- Poor quality
- Difficult voice design

Rule-based Synthesis – History

Resynthesis by Fant 1953

Resynthesis by Fant 1962

Fist complete TTS-system (Umeda 1968)

Klatt's TTS-system 1982

Acoustic Synthesis in Data-Driven Architectures

- Pre-recoded units do not fit the prosody of target utterance
- Necessary: Signal manipulation in the time domain (Fo and duration)
- If units are manipulated and concatenated, distortions at concatenation boundaries disturb quality
- PSOLA: spreads concatenation point across entire Fo period

PSOLA
pitch synchronous overlap add

- Elementary unit: interval of two weighted Fo-periods
- Consecutive intervals overlap each other
- Intervals are shifted and added appropriately
- Loss of quality if duration is stretched too long or Fo-manipulation more than half an octave
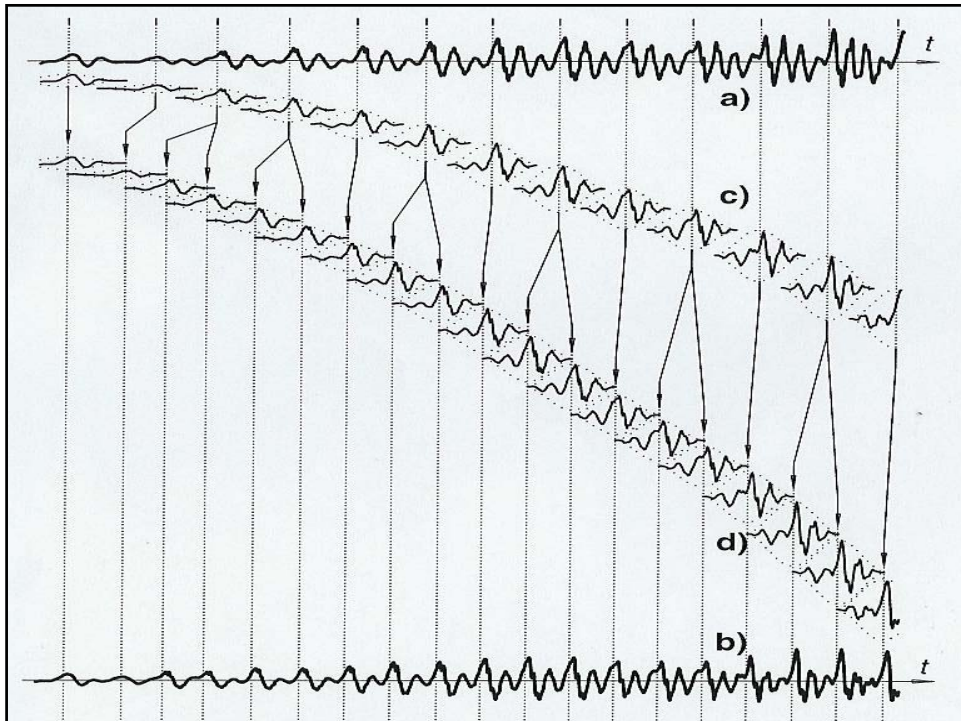
Corpus Construction in Data-driven Architectures

- Definition of unit inventory
- Carrier sentences, units in unstressed syllables

  „He has intere/ld/edee again"

- Careful recordings, several sessions (unsolved question: what's a good voice?)
- Avoid variation in speech rate, voice quality, intensity!
- Manual annotation ☹

Data-Driven Architectures – Pros and Cons

- Easy new voices („personal synthesiser" possible)
- Gain in naturalness, better synthetic quality
- Increase in quality facilitates research in functions of prosody (semantic, pragmatic)
- Prosodic manipulation limited
- Distance to articulatory model

Data-Driven Architectures – Examples

- Olive 1976, first system with concatenation of natural units
- Example for PSOLA-based system (ELAN)
- Diphone synthesis with very carefully recorded inventory (ETEX 2000)

Corpus-based synthesis – Progress or Capitulation?

- „State of the Art": Synthesis from Corpus
- In between „slot-and-filler"-systems and traditional concatenative systems
- Ideas:
  - „the best unit is the natural utterance"
  - Avoid manipulation by introducing more variants to units
  - „Chose the best to modify the least"

Unit Selection

Foreach matching synthesis unit in database (i.e. correct phone, phone sequence, word...)

    {
    Compare desired features
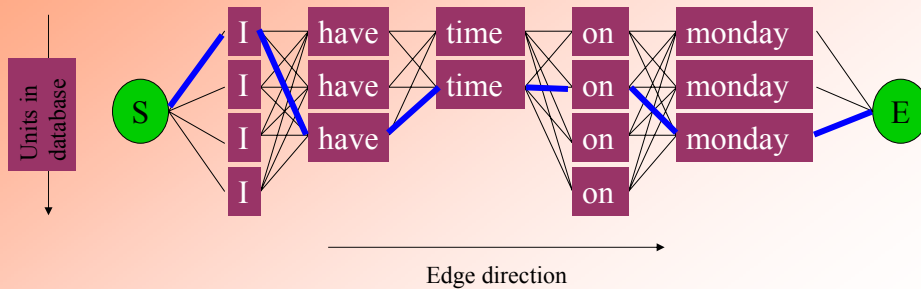   with
    unit features
    }

Determine optimal unit 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

Edge direction

New Architecture in Comparison

- Hadifix synthesis system (diphones, demisyllables, affixes)

- Corpus-based approach within domain

- Corpus-based approach out of domain

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...)

Nowadays...

- Most commercial architectures are based on unit selection synthesis
- systems: CHATR (traditional), NextGen (AT&T, „state of the art")
- Large corpora necessary – difficult to annotate for research institutes

but...

- Impossible to annotate/search all possible variations even for one speaker
- Therefore phonetic research still unavoidable
- Questions concerning appropriate speaking styles, emotions, even speech rate etc. remain unsolved

Evaluation of Synthetic Speech

- Diagnostic evaluation to localise systematic mistakes
- Global evaluation to assess overall
- Naturalness and intelligibility determine acceptability
- No natural reference makes auditory tests necessary

Evaluation of intelligibility

- Multiple choice („rhyme tests") or open response tests („CLID-test")
- Often phonotactically well-formed nonsense syllables, phonetically balanced
- Open response tests preferable to determine errors in unit inventory
- Syllable units inadequate to test consonant combinations across syllable boundaries
- Each subject can only be tested once

Evaluation of comprehension

- Did the subject comprehend the content of the synthesised text
- Phonetically balanced short texts are presented
- Subjects either transcribe the text or are asked questions concerning the content
- Subject able to characterise level of comprehension

Evaluation of naturalness

- Goal is not necessarily „perfect human voice" (sometimes this would create misunderstandings)
- Goal is to have synthetic speech as pleasant and easy to listen to as human voice
- Naturalness multidimensional
- Either preference tests or judgement scales
- Prosodic quality and voice quality both play crucial role in naturalness
- Delexicalisation used to test prosodic quality independent of segmental quality

Crucial Points in Evaluation

- Different applications may need different „types" of quality
- Different users may prefer different voices and different implementations (lively prosody or rather monotonic prosody, male or female voice, specific voice quality)
- Evaluation should be application specific

Where are we now?

- Working applications vs. Science Fiction (convincing emotions, the virtual actor...)
- System users usually do not like synthetic speech, corpus-based systems preferred immediately, rule based systems more stable in quality
- Alternatives? (Good old orthography, multi-modal domain-specific systems rather than unlimited domain TTS,...)

What I am interested in...

- Do you have any synthesis projects you are currently working on?
- Are there any projects you are working on where synthesis could be a method of evaluation, hypothesis testing etc.?
- Do you plan to use synthesis in some of your future projects?