

A Model for Prosodic Realization in Concept-to-Speech Synthesis

Klausurtagung of the Graduiertenkolleg
“Sprachliche Repräsentationen und ihre Interpretation”
Söllerhaus, Kleinwalsertal
Martin Haase
IMS, Universität Stuttgart

June 21, 2002

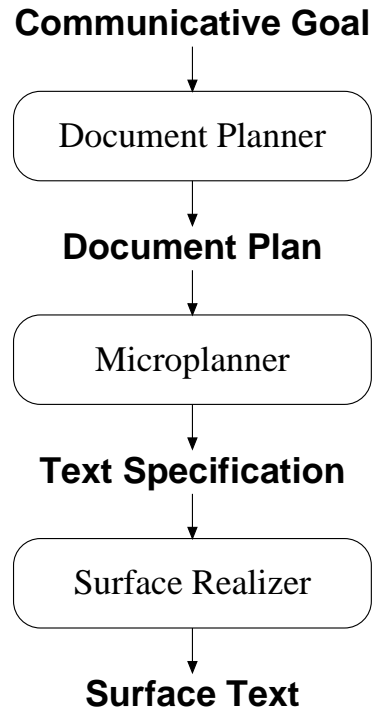
Speech synthesis and prosody

- Text-to-Speech (TTS) approach: *text analysis* process, followed by a *synthesis* process, cf. (Möbius, 2000)
- Output of analysis is a parameterization of the speech signal for waveform synthesis
- Many aspects of prosody must be specified, e.g.
 - intonation; place and type of pitch accents
 - prosodic phrasing
 - phrasal pitch range
 - segmental durations

Concept-to-Speech

- CTS - exploit 'linguistic' knowledge directly and omit analysis step
- Assume that source data are not in textual form
- Prevalent way to implement CTS: extend a natural language generation system to produce the parameterization of the speech signal, eventually accompanying the textual output
- Traditional field for CTS: spoken language frontend to a database, e.g. for information systems (Young & Fallside, 1979; Theune, Klabbbers, Odijk, & de Pijper, 1997; Teich, Hagen, Grote, & Bateman, 1997; Pan & McKeown, 1997))

Natural Language Generation: Questions

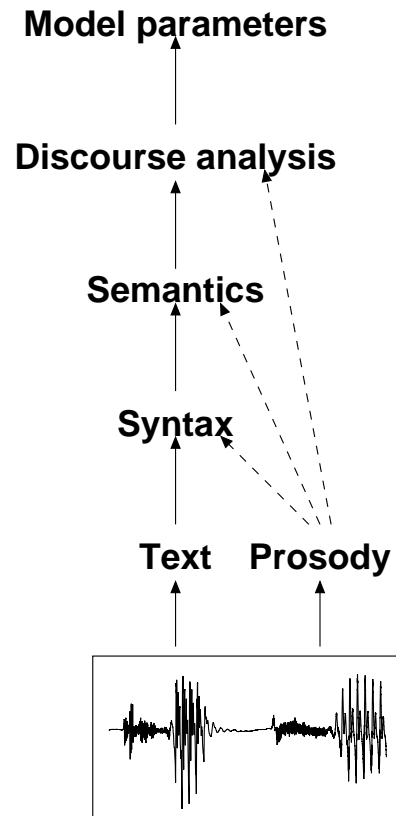


- What is the place of a ‘prosodic planner’ in current models of NLG, e.g. (Reiter & Dale, 2000)?
- How can planning decisions in a generation system be guided?
- Whereas the output of NLG/CTS is clear, what is the input?

Input representations

- Surface Realization
 - Finding a consensus for the input of NL generation amounts to finding a consensus for the output of NL analysis.
 - Models do exist: for specific theories, specific generation formalisms, grammars for specific domains (cf. (Reiter & Dale, 2000))
 - Restricted to single sentences
- Many prosodic phenomena only show up in the interaction of multiple sentences/phrases, but representations for generation above the sentence level are very much under debate
- Desirable approach: only specify input representations for CTS that are ‘grounded in reality’, e.g. D2S, (Theune et al., 1997); or (Conklin & McDonald, 1982)

Integrating prosodic planning



- Prosodic form of an utterance influences its syntax, semantics and its role in the discourse
- Consequently, prosodic realization decisions must be influenced by linguistic representations on multiple levels as well
- Motivate representations for NLG from research in linguistic analysis

Guidance of planning decisions

- Reversing a grammar? Ambiguity resolution vs. realization options
- Use models from language production → language performance rather than competence
- Language use as a process of constructing analogies with previously experienced utterances
- Analogy-based methods
 - Data-oriented Parsing: no explicitly formulated grammar, but use of an annotated corpus of parse trees (Scha, Bod, & Sima'an, 1999)
 - Application to NLG: instance-based natural language generation (Vargès, 2001)

Instance-based methods

- IB methods have been applied to various areas in computational linguistics (cf. (Daelemans, 1999))
- Common to analogy-based approaches is that they
 - use an instance-base (IB) of previously stored cases (e.g. an annotated corpus)
 - do not construct an abstract model ('grammar') but classify new data by comparing it to the cases stored in the IB
- Some advantages over classical statistical methods, e.g. need less data, better handling of 'low-frequency-events', but computationally demanding

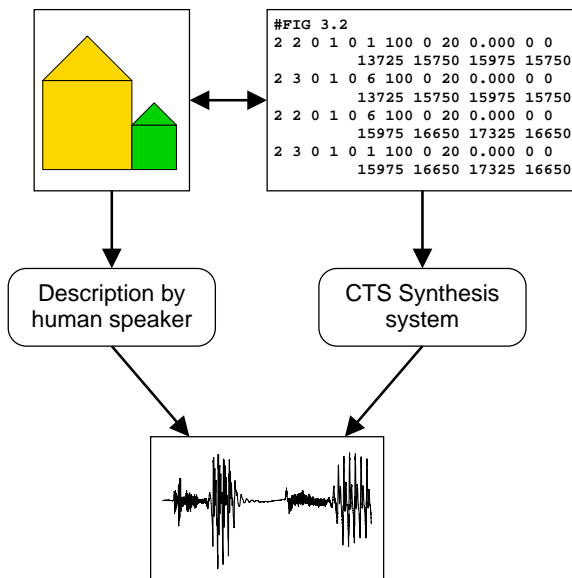
Speech production corpus

Guidelines for the project

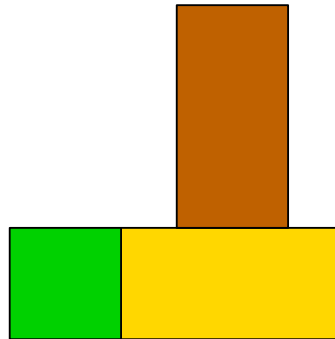
- Need an annotated corpus of spoken utterances together with the ‘conceptual representations’ *they are produced from*
- Multi-sentential utterances in spontaneous speech, not read speech
- Very restricted domain of corpus
- Ideally (...) large number of samples for learning / to use as IB
- Closest experimental paradigm (I found): Map task, or blocks world

Corpus design

- *Description* of simple figures (cf. also (Conklin & McDonald, 1982))
- Design experiment to enforce production of
 - Contrastive focus with respect to color, size, shape, and orientation; within and across figures
 - Multiple possibilities how to chain through complex figures, cf. (Levelt, 1982)
 - 'Figurative concepts', e.g. *house*, *arrow*, *cross* and aggregation of subparts
- Ask speaker to describe a figure as an annotation to a picture for a visually handicapped person



Example utterance

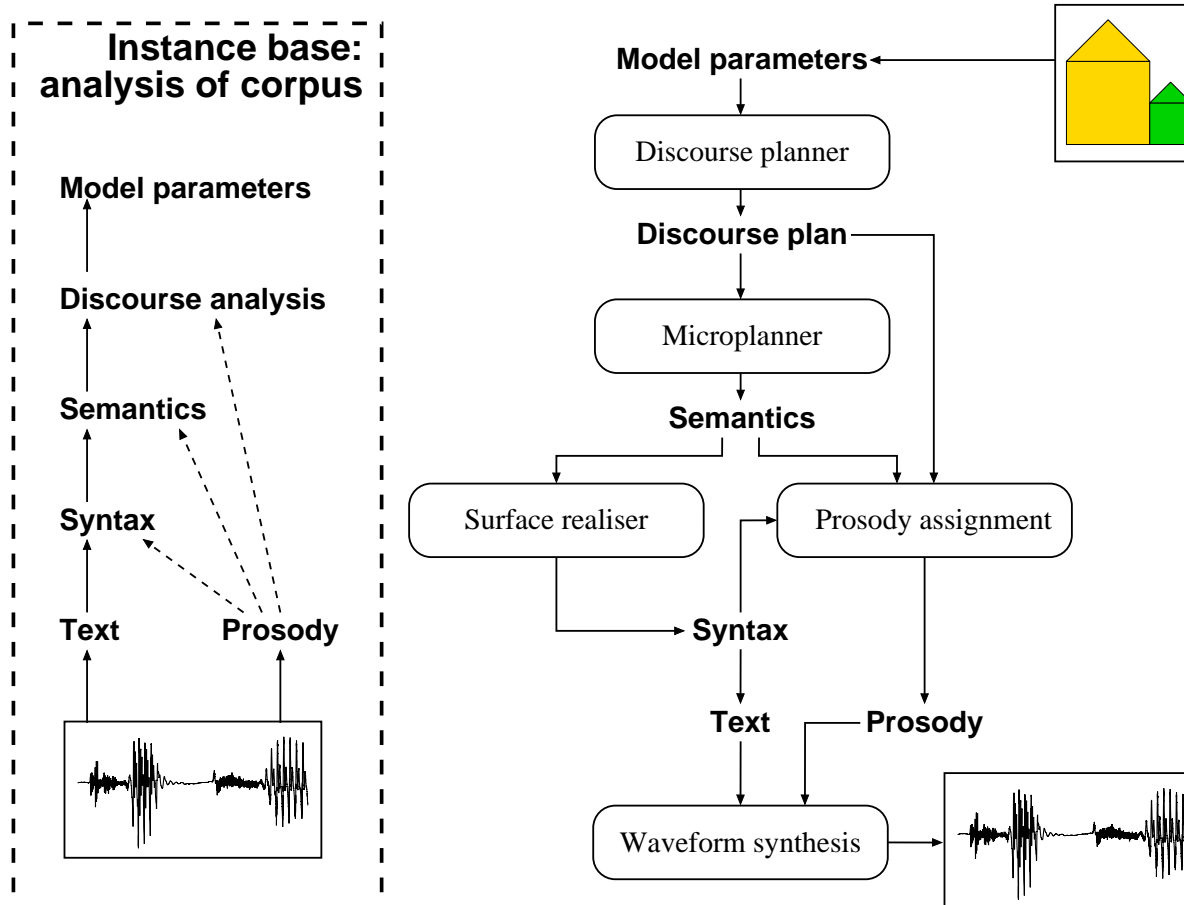


ein liegendes $\langle L^*H \rangle$ gelbes Rechteck $\langle L^*H \rangle \langle \% \rangle$
darauf $\langle *? \rangle$ stehend ein etwa so großes braunes $\langle L^*H \rangle$ Rechteck $\langle \% \rangle$
links $\langle L^*H \rangle$ liegend und anliegend am gelben $\langle L^*H \rangle$ Rechteck $\langle - \rangle$
ein grünes Quadrat $\langle H^*L? \rangle \langle \% \rangle$

Corpus construction

- Recorded 100 utterances of one speaker (10 blocks with 10 figures each) in the anechoic chamber of the phonetics group; ca. 30 min overall length incl. pauses
- Automatic analysis with tools available at the IMS: F_0 extraction, forced phoneme alignment, POS tagging. . .
- Human annotation: transliteration, prosodic labelling, syntax tree markup, semantic and discourse analysis (?)
- Extract geometric information from a figure's parameter file and code it in a PROLOG representation

Model for a CTS synthesis system



Questions

- What models can be used to align different levels of representation? (e.g. the NITE project, (Evert, Carletta, O'Donnell, & Vögele, 2002))
- What to do with disfluencies/hesitations?
- For which levels can IBL methods be applied? Use IBL and rule-based methods for different levels? Pure instance-based vs. case-based methods? Is the corpus large enough?
- Should *Concept-to-Speech* be replaced by *Data-to-Speech* (Theune et al., 1997)?

REFERENCES

*References

- Conklin, E. J. & McDonald, D. D. (1982). Salience: the key to the selection problem in natural language generation. In *Proc. 20th Ann Meeting of the ACL, Toronto*.
- Daelemans, W. (1999). Introduction to the special issue on memory-based language processing. *Journal of Experimental and Theoretical AI*, 11(3), 287–296.
- Evert, S., Carletta, J., O'Donnell, T. J., & Vögele, A. (2002). NITE corpus specification. Manuscript, IMS, Stuttgart.
- Levelt, W. J. M. (1982). Linearization in describing spatial networks. In Peters, S. & Saarinen, E. (Eds.), *Processes, beliefs, and questions*, pp. 199–220. Reidel, Dordrecht.
- Möbius, B. (2000). *German and Multilingual Speech Synthesis*, Vol. 7 (4) of *AIMS - Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung*. IMS, Universität Stuttgart. Habilitationsschrift.
- Pan, S. & McKeown, K. R. (1997). Integrating language generation with speech synthesis in a concept to speech system. In *Proc. ACL Workshop on Concept to Speech Generation Systems, Madrid*.
- Reiter, E. & Dale, R. (2000). *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press.

REFERENCES

REFERENCES

- Scha, R., Bod, R., & Sima'an, K. (1999). A memory-based model of syntactic analysis: data-oriented parsing. *Journal of Experimental and Theoretical AI*, 11(3), 409–440.
- Teich, E., Hagen, E., Grote, B., & Bateman, J. (1997). From communicative context to speech: integrating dialogue processing, speech production and natural language generation. *Speech Communication*, 21, 73–99.
- Theune, M., Klabbers, E., Odijk, J., & de Pijper, J. (1997). Computing prosodic properties in a data-to-speech system. In *Proc. ACL Workshop on Concept to Speech Generation Systems, Madrid*.
- Varges, S. (2001). Instance-based natural language generation. In *Proc. NAACL 2001*.
- Young, S. J. & Fallside, F. (1979). Speech synthesis from concept: a method for speech output from information systems. *J. Acoust. Soc. Am.*, 66(3), 685–695.

REFERENCES