A Cognitive Approach to Planning and Representation of Prosodic Features in a Concept-to-Speech System

Carsten Günther
Universität Hamburg
Fachbereich Informatik, AB Wissens- und Sprachverarbeitung (WSV)
Bodenstedtstr. 16, D-22765 Hamburg
e-mail: guenther@informatik.uni-hamburg.de

ABSTRACT

This paper presents a cognitive approach to prosodic planning in a language generation system. The macro- and microprosodic planning component is part of a computational modeling of main processing stages of the human language production process. The architecture of the system which is motivated by appropriate psycholinguistic insights, and some representational formats of prosodic knowledge are introduced.

INTRODUCTION

This paper is concerned with the phonological and phonetic planning component of the SYNPHONICS (Syntactic and Phonological Realization of Incrementally Generated Conceptual Structures) Formulator, and in particular with the representation of phonological and phonetic knowledge. The SYNPHONICS approach to the computational modeling of natural language production takes into consideration results from psycholinguistic research about the time course of the human language production process as well as recent developments in theoretical linguistics and phonetics concerning the representation of syntactic, phonological, and phonetic knowledge. The crucial point of linguistic investigation lies in the analysis and modeling of the syntactic and prosodic realization of different information structures (e.g. focus-background structure) in accordance with conceptual and contextual variations. The SYNPHONICS Formulator is the central part of the SYNPHONICS System[1], which is at present in a conceptual stage and will comprise the whole generative processing of utterances from pre-linguistic conceptual structures over complex semantic/syntactic/phonological structures onto acoustic parameter sets for controlling a speech synthesizer.

In our approach, language production is seen as an incremental process (Levelt 1989) which combines parallel and serial processing. Therefore, the planning processes must be hold local and must act over incomplete structures (e.g. there will be no preplanning of metrical trees or of complete intonation contours over whole utterances). This assumption about processing properties adheres to a special relational account of linguistic structures. In this account we assume that semantic, syntactic, and phonological information can be linked to each other, building a complex sign with inherent constraints. In abandoning a strictly functional dependency of phonological structure on syntactic structure we assume a direct interrelationship between semantic and phonological structure[2]. Such a view directly influences the organization of the prosodic planning processes and also the structure of the processing units (increments). In the next section, the architecture of the phonological and phonetic encoder within the SYNPHONICS Formulator will be described.

---

[1] For a detailed description of the SYNPHONICS system see Herweg (1992) and Schopp (1993).

[2] Evidence for the relevance of such a direct relationship between semantic and prosodic structure is shown by means of examples of focus/background structuring in Günther et al. (1993).

THE MODEL OF THE PHONOLOGICAL AND PHONETIC ENCODER

The architecture model of the phonological and phonetic encoder (Figure 1) with its processing steps of phonological, phonetic-articulatory, and acoustic encoding shows a clear separation of declarative knowledge components from procedural control components. This is due to the integration of a declarative grammar component (a variant of a HPSG (Pollard&Sag 1992) for German) in a procedural control structure. The grammar component is expanded by abstract semantics/phonology schemes (e.g., a Focus-Accent Scheme), a detailed phonological lexicon , and prosodic principles.

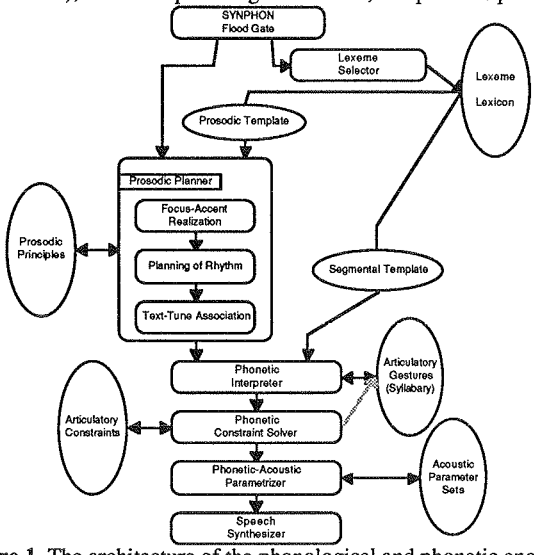

Figure 1. The architecture of the phonological and phonetic encoder.

Within the SYNPHONICS Formulator, we will represent phonological and phonetic information according to the event-based paradigm (Bird&Klein 1989). But the type of events and their properties will be separated into phonological and phonetic ones, reflecting the peculiarities of phonological and phonetic processes. During phonological encoding, we use a type hierarchy based on autosegmental phonology, whereas during phonetic encoding, we prefer a type hierarchy and property assortment that takes its bearings from articulatory processes (Browman and Goldstein 1989). A separation of phonological and phonetic encoding processes with adjusted structures increases the modularity of speech production models and applies to the research topic of current language generation investigation: Each decision or evaluation has to be carried out on its hereditary processing stage.

The *SYNPHON Flood Gate* forms the interface from semantic and syntactic to phonological planning. This module ensures the incremental subsequent treatment of already semantic and syntactic specified utterance fragments. The *Flood Gate* selects structure units which meet the inherent needs of phonological processes. According to psycholinguistic investigations (Levelt 1989), *phonological phrases* or *accent domains* (Gussenhoven 1983, Ladd 1983) - a semantic pendant to the well known syntactically defined phonological phrase - are conceivable to be such incremental units.

The *Lexeme Selector* selects the corresponding lexemes from the lexicon by dereferencing the lexeme pointer (an abstract address determined during lemma selection) and using syntactic agreement information as well as case information. Only this second

lexicon access (after lemma selection) makes available the concrete word form information (Levelt 1992a). In the *Lexeme Lexicon*, morphological, metrical and segmental information are stored. This information is specified during lexical-phonological spellout processes. Figure 2 shows a prosodic specified lexical entry of the proper name *Hans* in an HPSG-like style (semantic, syntactical, morphological, and concrete subsegmental event information is omitted).

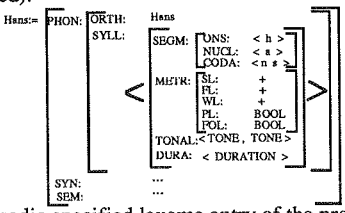

Figure 2. A prosodic specified lexeme entry of the proper name *Hans*.

The feature SYLLabic is a list of syllable templates with syllabic, metrical, tonal, and durational information. A complex type hierarchy of METRical types permits prominence rules to interpret feature values (of type BOOLean) as metrical grid positions. The number of grid levels is limited to five prosodic structure levels (syllable, foot, phonological word, accent domain, intonational phrase). The TONAL and DURAtional features are not specified till application of postlexical prosodic rules. The feature TONAL has two tone place holders because some syllables can carry an accent tone and a boundary tone (e.g., focused monosyllabic words at the end of a sentence).

The *Prosodic Planner* obtains results from the lexical-phonological spellout processes. The system's architecture allows a metrical and a tonal planning which is independent of the concrete segmental spellout. Such an architecture corresponds to the findings in psycholinguistic priming experiments of speech production (Levelt 1992b). Within this module, the macroprosodic planning takes place. The *Prosodic Planner* derives an abstract prosodic structure[3] of the utterance from semantic, syntactical and lexical information according to prosodic principles. Prosodic planning involves the projection of the focus structure onto prominence structure (specifying the feature METRic of the syllable template), the rhythmical planning (e.g., in order to avoid stress clashes) and the text-tune association (specifying the feature TONAL for representing pitch accent or boundary tone). The internal ordering of the prosodic planning steps has to follow prosodic structure building constraints, e.g. in accordance with mutual dependencies between metrical structures, boundary tones, and accent tones. Figure 3 shows an HPSG-like rule which ensures that the word accent bearing syllable of a focused word will become the most prominent one within an intonational phrase and carry a high accent tone.

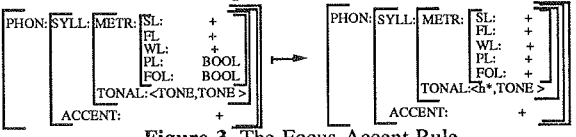

Figure 3. The Focus-Accent Rule.

The next processing stage is the *Phonetic Interpreter*, which forms our interface between phonology and phonetics and deduces a phonetic-articulatory event structure from abstract prosodic and segmental information by paying attention to segmental phonetic parameters. An interplay between global abstractly planned prosodic features and segment specific parameters takes place determining the concrete phonetic events which realize the prosodic

[3] The psychological reality of an abstract prosodic structure representation during sentence production was recently demonstrated by Ferreira (1993).

features. The standard articulator hierarchy of Browman and Goldstein's proposal (1989) is expanded by the articulator *jaw*, which is necessary in order to plan correctly the co-articulation effects and formant transitions of vowels. The phonetic interpretation of sub- and suprasegmental information relies upon a declarative knowledge base, the syllabary (Levelt 1992b). According to psycholinguistic investigations, the eventual increments which will be handed over from phonological encoding to phonetic interpretation are metrically structured, subsegmentally underspecified syllables or phonological words. These structures serve as the access code to the appropriate gestural score.[4] In order to model the cognitive language production process, we represent articulatory gestures as well as articulatory scores within our syllabary. Articulatory plans of syllables are already fully specified temporally and only the syllable environment has to be taken into account. But in case of assembling an articulatory unit from single gestures an articulatory constraint satisfaction process must be performed. At this point it is possible to implement a learning process which enlarges the syllabary in the case of frequently appearing articulatory scores.

The output of the phonetic constraint solver controls a speech synthesizer. Because of using an acoustic synthesizer, namely a Klatt-based formant synthesizer[5], a phonetic-acoustic interface is required. This module calculates the acoustic control parameters in accordance with the articulatory targets on the different articulatory event tiers.

ACKNOWLEDGMENTS

The research reported in this paper is carried out in a research project which is funded by the German Science Foundation (DFG) under grant no. Ha 1237/4-1.

REFERENCES

Bird,S. & E. Klein (1989): "*Phonological Events*", Research Paper EUCCS/RP-24. Centre for Cognitive Science, U Edinburgh.

Browman, C.P. and L. Goldstein (1989), "Articulatory gestures as phonological units", *Phonology*, Vol. 6, pp. 201-251.

Ferreira, F. (1993), "Creation of Prosody During Sentence Production", *Psychological Review*, Vol. 100, pp. 233-253.

Günther, C., Ch. Habel, C. Maienborn and A. Schopp (1992), "What's up with the printer? - Context relative presentation of conceptual structure and its prosodic realization in a language production system", in Schopp (1993), pp. 5-16.

Gussenhoven, C. (1983), "Focus, mode and the nucleus", *J. of Linguistics*, Vol. 19, pp. 377 – 417.

Herweg, M. (ed., 1992), *Hamburger Arbeitspapiere zur Sprachproduktion - I*, GK-Kognitionswissenschaft, AP 9, Uni Hamburg.

Ladd, D.R. (1983), "Even, focus and normal stress", *J. of Semantics*, Vol. 2, pp. 257-270.

Levelt, W.J. (1989), *Speaking: From Intention to Articulation,*. Cambr., Mass., MIT Press.

Levelt, W.J. (1992a), "Accessing word in speech production: Stages, processes and representations", *Cognition*, Vol. 42, pp. 1-22.

Levelt, W. J. (1992b), *Timing in Speech Production with Special Reference to Word Form Encoding*, Ms. MPI Nijmegen.

Pollard, C. & I. Sag (1992), *Head-driven Phrase Structure Grammar,* Ms., April 1992, Stanford, CSLI (to appear by CSLI).

Schopp, A. (ed., 1993), *Hamburg Working Papers on Language Production - II*, GK-Kognitionswissenschaft, AP 13, Uni Hamburg.

---

4 If articulatory scores are stored syllable-sized in the syllabary then the retrievel should reveal a frequency effect. And indeed, low-frequently syllables are harder to access as high-frequently ones (Levelt 1992b).

5 For synthesizing speech, a Klatt-based synthesizer was kindly made available to us by the Institute for Technical Acoustics of the Technical University of Dresden.