From Information Structure to Intonation: A Phonological Interface for Concept-to-Speech

Hannes Pirker, Georg Niklfeld, Johannes Matiasek
Harald Trost
{hannes,georgn,john,harald}@ai.univie.ac.at
Austrian Research Institute for Artificial Intelligence (OFAI)
Schotteng. 3, A-1010 Vienna, Austria

February 13, 1998

Abstract

The paper describes a component that interfaces between the generator and the synthesizer of a German language concept-to-speech system. It discusses phenomena in German intonation that depend on the interaction between grammatical dependencies (projection of information structure into syntax) and prosodic context (performance-related modifications to intonation patterns). The grammatical factors are covered by the unification-based generation grammar, whereas genuinely prosodic factors are implemented in the interface module, where influences like phonological distance between tonal accents are encoded more directly.

An extended two-level phonology component represents the core interface where the modules for grammar processing and speech synthesis meet and communicate. In a concept-to-speech system with its various modules built on diverse technological foundations, there is a strong case for having such a robust and flexible component that nevertheless offers a large degree of conceptual transparency.

As the overall objective of the project was to investigate whether and how conditions in concept-to-speech favour a more elaborate treatment of prosodic parameters in speech generation, a fairly complex model of phonology was required. Phonological processing in the system comprises segmental as well as suprasegmental dimensions such as syllabification, phenomena resulting in the modification of word stress positions, and a symbolic encoding of intonation contour. Phonological phenomena often touch upon more than one of these dimensions, so that mutual accessibility of the data structures on each dimension

had to be ensured. We present a linear representation of the multidimensional phonological data based on a straightforward linearization convention, which suffices to bring this conceptually multilinear data set under the scope of the well-known processing techniques for two-level morphology.

1 Introduction

The task of interfacing between a tactical generator and a speech synthesizer is two-fold: A grammatical description enriched with semantic and pragmatic features has to be translated into a phonological description. In a second step this (qualitative) phonological description has to be mapped onto the set of (quantitative) parameter values needed as input to the synthesizer. This paper will concentrate on the first task, a more detailed description of the second can be found in [Pirker et al. 97].

The requirements imposed by a concept-to-speech system differ from those on both text generation and text-to-speech systems (the most common application of speech synthesis). In text generation the generator proper produces a sequence of abstract descriptions of word forms which are–either by direct access to a lexicon or via a morphological component–transformed into strings of graphemes and output.[1] With concept-to-speech systems the task is more complex. Not only is segmental information influenced by morphonology and post-lexical rules (covering, e.g., reduction and assimilation phenomena) but–more important–suprasegmental information must be provided as well.

Compared to text-to-speech the task is at the same time easier and more diffcult. Easier, because information from pragmatic, semantic and syntactic layers are readily available. This eliminates the need to analyze an input text for the necessary cues to come up with proper pronunciation and prosody. More difficult, because all this information must be properly accounted for to come up with an adequate description of the utterance that–when fed into the synthesizer–produces high-quality output. This implies in particular that pragmatic-semantic features must be mapped onto (abstract) prosodic features.

In our system we employ finite-state morphology techniques, namely an extended version of two-level morphology for this interface.[2] The great ma-

[1] We leave aside problems like formatting and lay-out which are usually dealt with separately if at all.

[2] The extension regards the fact that the system allows the use of (feature-based) exter-

jority of applications for two-level morphology is in text processing. There exist a few exceptions, e.g., systems that translate letters to "sounds" for text-to-speech synthesis ([Williams 94], [Russi 92]) but we are not aware of any other application that deals purely with phonemic descriptions.

Nevertheless, the formalism proved to be very well suited for the task. The various almost independent subsystems can be kept conceptually separate resulting in good transparency while at the same time enabling the necessary amount of interaction between them.

2 A Concept-to-Speech Generation System

The concept-to-speech generation system consists of a pipeline of modules (cf. Fig. 1). A text planning component produces sentence plans, which are fed into the tactical generator. The tactical generator has two layers. The first one is dealing with sentence level generation, producing a tree-like description of a sentence, the leaves of which are lemmata annotated with morphosyntactic and prosodic features. The second layer performs generation at the word level producing annotated phonological representations of the correctly inflected word forms. These representations are fed into the extended two-level phonology component applying morphological and phonological rules to arrive at the representation used as input for the speech synthesizer.

Determining and encoding prosody mainly involves the tactical generator and the two-level component.

2.1 The Tactical Generator

The implementation basis for the tactical generator is the FUF [Elhadad 91] system. FUF is based on the theory of functional unification grammar and employs both phrase structure rules (being encoded by means of a special CATegory feature) and unification of feature descriptions. Input to FUF is a partially specified feature description which constrains the utterance to be generated. Output of FUF is a fully specified feature description (in the sense of the particular grammar) subsumed by the input structure, which is then linearized to yield a sentence.

---

nal information to restrict the application of two-level rules. For a description of two-level morphology in general, see [Karttunen & Beesley 92], for extended two-level morphology, see [Trost 91].

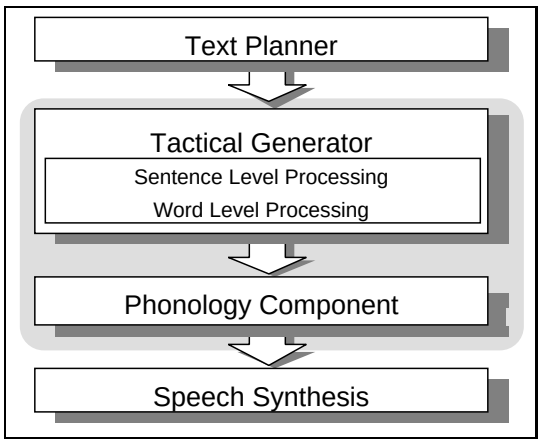

Figure 1: Architecture

A distinguishing feature of the grammar used in the generator is the integration of sentence-level and word-level processing within the same formalism, facilitating the information transfer between these two levels.

2.2 Extended Two-Level Processing

The two-level formalism employs the extensions proposed in [Trost 91], i.e., two-level rules may be annotated with filters. An annotated rule applies only, if its filter unifies with the feature structure the current string is annotated with. The filter handling uses the FUF formalism and the same unification machinery as the grammar.

This architecture forms an ideal platform for the implementation of the phonological interface. The necessary adaptions are limited to the data used: An existing grammar was extended with features describing the information structure. The lexicon consists of entries in phonemic form (using SAMPA notation) enriched with information like (potential) accent and syllable boundary positions. Scope and coverage of the two-level rules are described in section 3.

2.3 Speech Synthesis

Input to the synthesizer is a SAMPA string enriched with qualitative encodings of prosodic information (e.g., pitch accent, pauses, ...) produced by the two-level rules.

The phonological specifications of intonation are processed by a phonetic interpreter [Pirker et al. 97], that performs the transformation of these qualitative labels into quantitative acoustic parameters. The synthesizer then picks the right demisyllables from the inventory, smoothes the transitions and modifies pitch contour and duration according to the parameter settings.

Although some interpretative work is done within the synthesizer, no linguistically motivated transformations are supposed to take place there. These all are performed within the two-level component.

3 The Phonological Interface

3.1 Phenomena handled

The phonological description in extended two-level morphology – in our case rather two-level *phonology* –serves as the central interface where the modules for grammar processing and for speech synthesis meet and communicate.

A fairly complex model of phonology is required in the system, also because the overall objective of the project was to investigate whether and how conditions in the concept-to-speech task favour a more elaborate treatment of prosodic parameters in speech generation.

The phonological description is implemented in the extended two-level framework described in section 2.2 and works over a lexicon of phonemic (rather than graphemic) representations of word stems and inflectional affixes. Morphotactic processing is thus restricted to inflection, whereas compounding and derivational affixation are encoded in the lexicon, which is typically small in domain-tailored concept-to-speech systems.

Nevertheless, in segmental phonology, the component must compute morphonological rules in inflection as well as post-lexical rules which interact with syllabification and cliticization. Well-known rules of this type in German phonology are for example umlaut and final devoicing, which are both covered by the component.

To determine German syllabification and cliticization correctly, it is necessary for the phonology to operate on structures larger than single words [Laeufer 85]. Consider for example phonological cliticization, i.e. resyllabification on the phrase level which constructs syllables that stretch across word boundaries. This phenomenon is heavily restricted in German (unlike in French), but resyllabification nevertheless occurs even at slow-to-normal speech rates, namely with unstressed personal pronouns that start with a vowel. Such pronouns are syllabified together with the preceding word.

In order to cover such phenomena, phonological processing in the component applies to chunks whose size depends on the one phonological rule in the system that requires the largest phonological context to operate correctly. Because of the intonation rules discussed below in section 4, phonological processing applies to the whole utterance[3].

It may be worth mentioning that the three phonological aspects of segmental representation, syllabification, and word stress are mutually dependent in German phonology in all logically possible directions
[Niklfeld et al. 95]. The phonology component treats them in a unified description, which also covers the rare cases of word-internal and phrase-level stress shift in German.[4] To give a flavour of the data, stress shift can for instance apply when the nominal head of a direct object which has ultimate stress immediately precedes a particle verb with stressed initial particle. E.g. *Hút aufsètzen* ("hat on-put", to put a hat on), with stress shift applied to the verb.

While some segmental and supra-segmental rules in the phonological description depend on phonological context only, some others (like the rule for stress shifts as described above) depend on grammatical information on levels as high up as textual representation. For example, the German word for "weather" loses word stress in compounds when they appear in weather-reports (where the concept weather is "textually exophoric" [Benware 87]). Such phenomena are encoded in our extended two-level system by phonological rules which access the grammatical representation via feature-filters.

There are few theoretical frameworks in computational linguistics for tackling such a breadth of phonological issues. Linguistically ambitious approaches are often designed with little regard to ease of use in large descriptions, whereas leaner formalisms do not scale well to complex data stretching across a number of phonological dimensions. The chosen framework of extended two-level phonology stands between these poles.

3.2 Linearization of multi-tier phonological structures

As the two-level framework assumes one lexical and one surface string only, we use a linear representation of our multidimensional phonological data, as follows:

Each linear phonological string in the component stands for a multi-tier structure which combines a given number of separate dimensions of

[3] I.e., syllables, phonological words, intermediate phrase, intonational phrases – the description does not make use of feet.

[4] Otherwise, German has lexically specified word stress.

phonological structure. The tier of phonological segments (members of the German SAMPA set) is used to provide the backbone of skeletal points on which all units of the representation are linked together. Each unit on any phonological tier has scope over/has as its domain a continuous section of skeleton points. For each tier, a convention is provided which designates that part of each domain which is used for the linking. For some suprasegmental tiers (syllables, phonological words) the leftmost unit of the scope domain is used for this purpose. For other tiers the domain edges are not specified in the lexicon (stresses and accents, which have scope over stretches of syllables), and therefore other well-defined parts of the scope domain are used for the linking (such as the vocalic nucleus of a syllable). Where it appears natural to do so, units on certain phonological tiers are also linked to right domain edges (as is the case with phrase and boundary tone markers, which have scope over any phonological material between a nuclear tone and the right boundary of an intonation phrase.)

While these representations clearly encode some fragment of autosegmental phonology in an implicit way, they do not allow for the attachment of more than one suprasegmental unit from the same tier to a single segmental unit. This power was not needed in our application, and the described phonological representation proved a handy way to conceptualize the phonological phenomena we are dealing with. It also allowed for easy incremental extensions to our descriptions, as additional tiers of representation could be added as coverage of higher-level prosodic issues such as sentence intonation improved.

3.3 Implementational notes

Using the linearized representation, the well-known processing schemes for two-level morphology [Karttunen & Beesley 92] can be applied directly. Contemporary compilers for two-level morphology allow to specify sets of symbols that are ignored in individual rules. Extensive application of such syntactic sugar enables us to keep the rule formulations over the collapsed representation economical and relatively transparent. We note in passing that although collapsing multilinear data-structures onto a single tier increases the likeliness of combinatorial explosion in processing when using the two-level automata as transducers, it turns out that in our already quite complex description this does not become a real problem.

In earlier publications, we described how we implement phonological generalizations that stretch across phonological dimensions [Niklfeld et al. 95], and we proposed implementations of suprasegmental issues such as stress

shift and the projection of pitch accents depending on focus information [Niklfeld & Alter 96]. We have also discussed time structure [Alter et al. 96]. In section 4 we go beyond this to show that intonation in German has properties that are best implemented by combining our two-level phonological description, which is well-suited to express constraints on linear contexts, with the power of a unification-based feature grammar.

4 Dealing with Intonation

This section describes, how the extended two-level component is used for dealing with the problem of specifying "appropriate" intonation contours and phrasing.

4.1 Different perspectives

Intonation is influenced by a diverse collection of factors, e.g., discourse structure, pragmatics, semantics, syntax and phonology. Resesarch that is concerned with intonation (or prosody as a whole) thus displays a big variety of perspectives on that topic.

Inspecting the form. Amongst phonologist and phoneticians the autosegmental tone sequence model of [Pierrehumbert 80] has become the prevalent theoretical framework for "talking about intonation". In this framework intonational tunes are composed of sequences of H(igh) and L(ow) tones. Tones are classified into two groups namely *pitch accents* (which have a prominence lending function) and *edge tones* (which signal phrasal boundaries).

In 4.1 the inventory of the GToBI system [Grice et al. 96] is displayed, which also forms the basis for our phonological concept. GToBI is a German variant of the original ToBI (*Tones and Break Indices*) presented by [Beckman & Ayers 94].

Don't bother about form. On the other hand there is a rich tradition of research, that may be subsumed under the rather abstract keywords of "accent placement". The overall orientation of these works may be either rather syntactically (keyword: focus projection), semantically or pragmatically (keyword: given vs. new information) oriented. A common feature of these approaches is the omission of phonological aspects, i.e., they deal with the *distribution* of accents and not with their *form*.

Form follows function. Another strand of research deals with the coupling of semantics (or more specifically information structure) and phonology, i.e., the association of meanings and tunes. For instance [Hobbs 90]

H*	'peak accent'
L*	'low accent'
L+H*	steep rise
L*+H	Rise. Peak moved behind accented syllable
H+L*	Step down to low range
H+!H*	Step down to mid range
L- H-	Phrase tones
L% H%	Boundary tones

Table 1: The tonal inventory of GToBI

associates (trailing) H-tones with the notion of *incompleteness* and *openendedness*, an assertion that fits well into the commonly observed predominance of rising contours in questions.

In [Prevost & Steedman 94] is a rare example for the tight coupling of information structure and intonational phonology implemented in a speech system is presented. The classification of the utterance's elements along the dimensions *theme/rheme* and *focus/ground* trigger the selection of tones in a straighforward fasion: focused concepts in the theme, e.g., will uniformly be assigned with a rising L+H* while focused parts in the rheme receive basically falling accents.

Form follows heuristics. An approach rather common in the field of text-to-speech synthesis is the assignment of markers for prominence and phrasing on the basis of algorithms and heuristics that intermingle information on syntax, punctuation, word-class information as well as phonology and phonetics in a rather unstructured way.

4.2 Our design

In our system we employ a strict separation: only the two-level component deals with tonal specifications. Within the tactical generator only the candidate *positions* for both pitch accents and phrasal boundaries are selected.

This reflects the fact that though prosody depends heavily on grammatical and pragmatic factors, its definite realisation is also strongly influenced by phonological and phonetic constraints which are much more "naturally" handled by the two-level component than by the generator. Expressed in the terminology of two-level morphology the grammar provides a underspecified "lexical" representation from which the concrete "surface" form is derived.

In the lexicon every (accentable) word contains an abstract pitch tone (T) within its phonemic representation. The "lexical boundaries" (B), i.e., candidates for boundaries between intonational phrases (IP) are inserted by the generator in between words.[5] The two-level component then either maps it to a GToBI label or discards it i.e., maps it to surface 0.

The following example (presented in pseudo-code) defines a basic condition on the IP: it contains at most three pitch accents, at least one, and has an obligatory boundary tone.

```
<IP>       ::= {<PitchTone>{<PitchTone>}}
                <PitchTone><IP_Boundary>
<IP_Bound> ::= L-L% | L-H% | H-L% | H-H%
<PitchTone>::= <RisingT> | <FallingT>
<RisingT>  ::= H* | L+H* | L*+H
<FallingT> ::= L* | H+L* | H+!H*
```

In order to determine the realisation of a T the grammatical information the generator provided for the word in question is inspected via the filter mechanism. In the simplest case this reduces to a lookup on whether the grammar has marked the word as unaccented (acc -). In this case the tone will be ignored.

```
T:0 <= _ filter:
        (head (phon (acc -)));
```

In our simplified grammar L-L% and H-H% are exclusively used at the end of assertive sentences and questions respectively [6]. While all the rules discussed so far have been pure filter applications the last rule encodes a constraint on phonological context.

```
B:L-L% <=> _ filter:
          (head (s-type assertive));
B:H-H% <=> _ filter:
          (head (s-type interrog));
B:L-H% => <FallingT> <UnaccSyll>* _ |
          <RisingT> <UnaccSyll> <UnaccSyll>+ _;
```

[5] For the sake of simplicity we will neglect intermediate phrases here.

[6] In this architecture boundaries are just treated like ordinary words by the generator. I.e., they have an associated "lexical entry" where the utterance type is stored and thus can be inspected by the filter. Of course the use of additional unambiguous "lexical" boundary marks (e.g. using a Q at the end of questions that always becomes surface H-H%) would be possible as well, and of course more efficient. Nevertheless we decided to keep tonal features strictly out of the grammar.

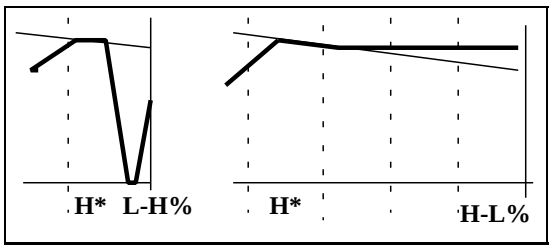

Figure 2: Contours to be avoided (vertical lines designate syllable boundaries)

The rationale behind this rule is, that we want to avoid the contours shown in figure 2 when realising IP boundaries. The `L-H%` boundary basically designates a fall-rise contour. This should be a felicitous contour if the last pitch accent before the boundary was a falling one. The second term states, that after a rising pitch accent the same boundary contour is to be produced only if the pitch peak is followed by two or more unaccented syllables thus ensuring that there is "enough room" for the fall rise. At the same time the production of the concurring `H-H%` is blocked, which would produce a monotonuous stretch on a high level, that might be perceived as unnatural.

The rules also implement some of the variability in prosody that is due to the interaction of phrasing and pitch accents much in the spirit of tone-linking [Gussenhoven 84].

5 Conclusion

In comparison to the the architectures outlined in 4.1 the following points may be emphasized. The handling of accentuation and phrasing by the generator resembles the syntacto-semantic approaches. We use only a few tags such as emphasis `[EMPH]` and (conceptual or textual) givenness `[GIVEN]` which are rather easily identifiable by the conceptual component and have a straightforward influence on the phonetic realisation. In this respect our approach is less refined than, e.g., [Prevost & Steedman 94] as no fully fledged semantic module is integrated that could deal with aspects of information structure in a really principled way

On the other hand we employ a very flexible and transparent phonogical model. Using GToBI, we rely on the wealth of phonological research undertaken in the tone sequence paradigm. But not all intonation contours that can be observed in human speakers are equally convenient for the use

in synthetic speech, where the deviations in duration, amplitude, etc. may lead to results that are perceived as highly unnatural. We thus restrict the set of possible contours licensed by the GToBI to a simplified subset.

The system is implemented and deals with the task of generating monologuous weather reports.

Acknowledgements

The work reported here has been carried out within the project *Phonology-Acoustics-Conversion for Concept-to-Speech (FWF P10822)* funded by the Austrian Fonds zur Fẗderung der wissenschaftlichen Forschung. Financial support for the Austrian Research Institute for Artificial Intelligence is provided by the Austrian Federal Ministry of Science and Transport.

References

[Alter et al. 96] Alter K., Matiasek J., Niklfeld G.: Modeling Prosody in a German Concept-to- Speech System, in Gibbon D.(ed.), Natural Language Processing and Speech Technology, Mouton de Gruyter, Berlin, 1996.

[Benware 87] Benware W.A.: Accent Variation in German Nominal Compounds of the Type (A (BC)), Linguistische Berichte, 108 pp.102-127, 1987.

[Beckman & Ayers 94] Beckman M.E., Ayers G.M.: Guidelines for ToBI Labelling, Ohio State University, Columbus, OH, 1994.

[Elhadad 91] Elhadad M.: FUF: The Universal Unifier User Manual, Dept.of Computer Science, Columbia University, 1991.

[Grice et al. 96] Grice M., Reyelt M., Benzmüller R., Mayer J., Batliner A.: Consistency in Transcription and Labelling of German Intonation with GToBI, Proc. of ICSLP 96, Philadelphia, pp.1716-1719, 1996.

[Gussenhoven 84] Gussenhoven C.: On the grammar and semantics of sentence accents, Dordrecht: Foris, 1984.

[Hobbs 90] Hobbs J.R.: The Pierrehumbert-Hirschberg theory of intonational meaning made simple: comments on Pierrehumbert and Hirschberg, in Cohen P.R., et al.(eds.), Intentions in Communication, MIT Press, Cambridge, MA, pp.313-323, 1990.

[Karttunen & Beesley 92] Karttunen L., Beesley K.R.: Two-Level Rule Compiler, XEROX PARC, Palo Alto, CA, Technical Report ISTL-92-2, [P92-000149], 1992.

[Laeufer 85] Laeufer C.: Some Language-Specific and Universal Aspects of Syllable Structure and Syllabification: Evidence from French and German, PhD-Thesis, Cornell University, 1985.

[Niklfeld et al. 95] Niklfeld G., Pirker H., Trost H.: Using Two-Level Morphology as a Generator- Synthesizer Interface in Concept-to-Speech, in Proceedings of the 4th European Conference on Speech Communication and Technology, Madrid, Spain, Vol.2,pp.1223-26, 1995.

[Niklfeld & Alter 96] Niklfeld G., Alter K.: Covering prosody in concept-to-speech via an extended two-level-phonology component, in Computational Phonology in Speech Technology - Second Meeting of the ACL Special Interest Group in Computational Phonology, University of California, Santa Cruz, CA, Association for Computational Linguistics, 1996.

[Matiasek & Trost 96] Matiasek J., Trost H.: An HPSG-Based Generator for German - An Experiment in the Reusability of Linguistic Resources, in *Proceedings of the 16th International Conference on Computational Linguistics, August 5-9, Copenhagen, Denmark*, Center for Sprogteknologi, Copenhagen, Denmark, pp.752-757, 1996.

[Pierrehumbert 80] Pierrehumbert J.B.: The Phonology and Phonetics of English Intonation, Ph.D. Thesis, MIT, Cambridge, 1980.

[Pirker et al. 97] Pirker H., Alter K., Matiasek J., Trost H., Kubin G.: A System of Stylized Intonation Contours for German, in 5th European Conference on Speech Communication and Technology, University of Patras, Greece, Vol.1, pp.307-310, 1997.

[Prevost & Steedman 94] Prevost S., Steedman M.: Specifying Intonation from Context for Speech Synthesis, Speech Communication, 15:139-153, 1994.

[Russi 92] Russi T.: A framework for morphological and syntactic analysis and its application in a text-to-speech system for German, in Bailly G. & Benoit C.(eds.), Talking Machines, Amsterdam, New York, Oxford: North-Holland, pp.163 -182, 1992.

[Trost 91] Trost, H.: X2MORF: A Morphological Component Based on Augmented Two-Level Morphology, in: IJCAI-91, Morgan Kaufmann, San Mateo, CA, pp.1024-1030, 1991.

[Williams 94] Williams, B.: Welsh letter-to-sound rules: Rewrite rules and two-level rules compared. Computer Speech and Language, vol. 8 (1994): 261-277.