Modeling Prosody in a German Concept-to-Speech System*

Kai Alter, Johannes Matiasek, Georg Niklfeld

Abstract

We present a concept-to-speech system for German which allows for a correct processing of prosody at the sentence level. Prosodically relevant features are derived from the informational (discourse) structure. These features comprise not only tone structure, but also time patterns related to focus-background-division, which have turned out to play an important role in synthesis as well.

Wir stellen ein Concept-to-Speech System für das Deutsche vor, welches eine korrekte Verarbeitung satzprosodischer Eigenschaften zu gewährleisten vermag. Die für die Prosodie wichtigen Eigenschaften werden der linguistische Struktur, insbesondere der Informationsstrukturierung, entnommen. Zusätzlich zur tonalen Struktur werden zeitstrukturelle Eigenschaften, die in Beziehung zur Fokus-Hintergrund-Gliederung stehen, ebenfalls bei der Sprachsynthese berücksichtigt.

1 Introduction

In this paper, we present work on the implementation of the correspondences between focus structure and its syntactic and phonetic correlates in a concept-to-speech system for German. First, based on empirical investigations of speech signals, an analysis of the main mechanisms for expressing focus information in the German linguistic system is developed. Implementing this account, the processing of focus information in tactical generation is described. The discussion of data is restricted to only a small sample of the domain of linguistic focus in German; rather than descriptive breadth, the goal here is to present an implemented account that spans all the different linguistic levels involved in the expression of focus: From the determination of focus information on the discourse-level, through syntactic phenomena of word-order variation, prosodic phonology, down to the correlated phonetic phenomena.

*This work has been sponsored by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung under grant no. P10822-MAT. Financial support for the Austrian Research Institute for Artificial Intelligence is provided by the Austrian Federal Ministry of Science, Transport, and the Arts.

2 Observations on the tonal and temporal structure of German sentences

In general, prosodic phenomena depend on at least 3 levels of sentence production:

1. grammatical information about
   - discourse structure (focus-background division, topic-comment division)
   - word order
2. restructuring of syntactic constituent-structure into larger phonological constituents (phonological phrases and higher levels);
3. performance structure (speech rate, speech style, complexity of constituents etc.).

In this paper we will concentrate on the first point and try to elaborate in which way discourse structure and word order influence the tonal and temporal structure of audible speech.

2.1 Tonal structure

Following (Pierrehumbert, 1980), pitch accents are derived from abstract focus features in the feature grammar. The corresponding tone features are assigned in the phonological representation.

In accordance with other work on focus (Dogil, 1995; Günther, Maienborn, and Schopp, 1994) we distinguish 3 types of focus: Wide focus (= sentence or normal focus), narrow focus and contrastive focus. For this presentation we only deal with wide and narrow focus. Contrastive focus may to a great extent be viewed as a special case of narrow focus from a prosodic point of view (e.g., the case of complementary contrast). Generally, in the following we discuss only a small part of the linguistic focus system in German, our emphasis here being on key features of the problem rather than on comprehensive discussion of the data.

The focus features discussed below are part of the unification-based representation common to all layers of linguistic description. Although these features may influence other aspects of generation as well (e.g., word order) they mainly affect prosody and our discussion will focus on these effects.

`(f-type wide)` is assigned with scope over the whole sentence under certain circumstances in text generation. In dialogs these circumstances can be forced by asking "What is happening?".

(1) [Die Sonne hat den Nebel vertrieben.]$_{\text{wide focus}}$
the sun has the fog driven away
"the sun has driven away the fog"

In our system this feature is assigned every time the discourse moves on to a new topic. Fig. 1 shows the intonation contour for this sentence.[1]

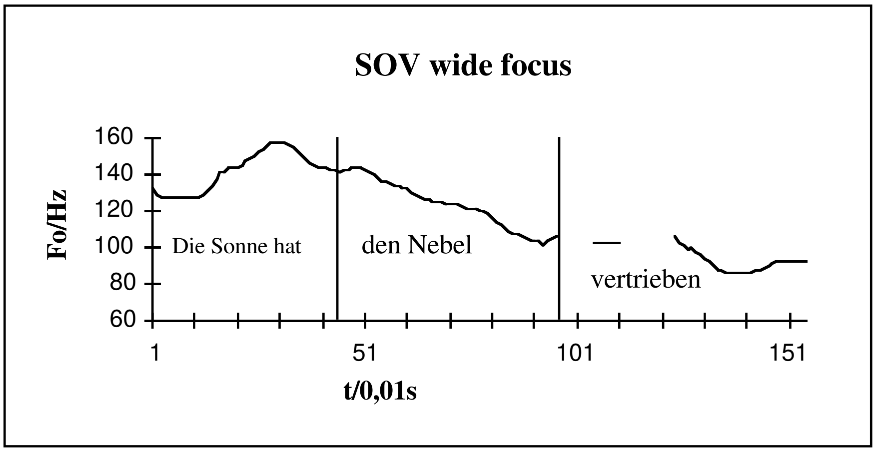

Figure 1: Wide focus, sentence accent on the object "den Nebel" (the fog)

`(f-type narrow)` is assigned to a single constituent which is in the focus of the sentence, i.e., to a part of a sentence which is new for the hearer from the speaker's view. Here we consider only narrow focus on the direct object. As an example consider the answer to the direct question "What has the sun driven away?" (see fig. 2):

(2) Die Sonne hat [den Nebel]$_{\text{narrow focus}}$ vertrieben.
the sun has the fog driven away
"the sun has driven away the fog"

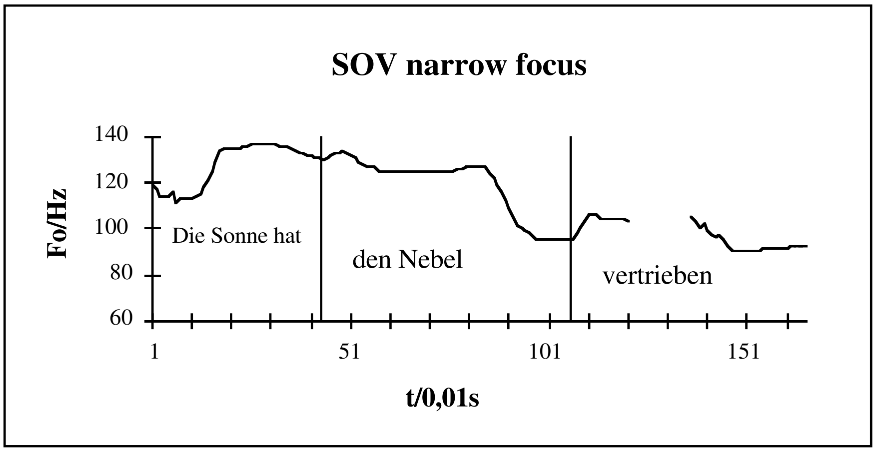

Figure 2: Narrow focus on the object "den Nebel"

[1] This and the following contours are part of a larger corpus we have collected from the speaker of our demi-syllable inventory.

The different bracketing in (1) and (2) result from different communicative intentions according to the discourse situation and originates from the text planning component. It can be used not only for deriving tonal features but also for the realisation of appropriate time patterns (see below).

We are now in a position to discuss two of the ways in which focus features interact with suprasegmental properties.

- In both sentences, our speaker uses two accents: A prenuclear accent on the subject and a focal nuclear accent on the argument. In each case, the accent is realized on the metrically strongest position of the constituent.
- The shape of the melodic excursion at the accent positions is dependent on focus type. We distinguish only two tonal structures of pitch accents for German phonology, a High-Low (H*L) and a Low-High (LH*) pitch accent[2].

We make two claims concerning the choice of tonal structures for the realization of the two pitch accent positions in the structures under examination:

First, nuclear accent assigned under wide focus to an argument or local adjunct immediately preceding the non-finite verbform is realized as a H*L, but nuclear accent expressing narrow focus on the constituent in the same position is realized as LH*.

Secondly we consider the prominence on the subject, or "topic" position. There is an alternative word order available if we have narrow focus on the direct object, which we refer to as topicalized[3] OSV (fig. 4):

(3) [Den Nebel]$_{\text{narrow focus}}$ hat die Sonne vertrieben.
the fog has the sun driven away
"the sun has driven away the fog"

Pitch accent on the topicalized direct object is triggered grammatically here by the presence of narrow focus, whereas the pitch accent on the subject which occupies topic position in the previous examples exhibiting neutral word order is generally hypothesized to be optional, i.e., triggered by performance. In tonal structure, however, we see no difference here and assign LH* to the topological topic position, independent of the focus type applying in the sentence.

2.2 Temporal structure

A second important way how focus can be expressed in German is by means of time structure. In an analysis of a question-answer-corpus collected especially for that purpose we find differences in the duration of constituents depending on

[2] There are various different labelling systems for German prosody, each having its own strengths. Our labelling system is optimized towards the internal architecture of our generator and the particular characteristics of the text genre (Garner, 1995) to be produced and combines features of the Stuttgart and the SYNPHONICS system.

[3] 'Topicalization' means the syntactic notion of a topic, e.g. the overt movement of a non-subject-DP in a sentence initial position. This notion does not correlate with the semantic concept of topic discussed in., e.g., (Bühring, 1995).

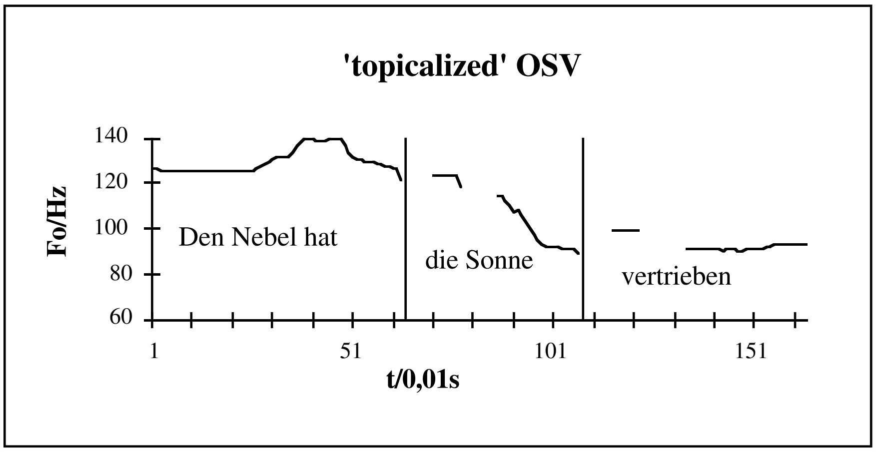

Figure 3: Focus on the topicalized object "den Nebel"

their position and on focus-background division. According to recent phonological hypotheses, phrasing in German is sensitive to focus (Günther, Maienborn, and Schopp, 1994; Selkirk, 1995). Phrasing and time patterns are correlated, in other words, phrasing may be realized phonetically by means of boundary signals (Alter, 1996). The following phonetic patterns may occur as boundary markers (BM):

(BM-1)	boundary tones
(BM-2)	pause marking
(BM-2.1)	final lengthening
(BM-2.2)	pause insertion
(BM-2.2.1)	glottal stop
(BM-2.2.2)	laryngalization
(BM-3)	lengthening in stress positions

However, the analysis of our corpus gave a surprising result: only (BM-3), i.e., lengthening in stress positions, was used for the marking of boundaries. Furthermore, (BM-3) has to be reinterpreted in terms of its phonetic realization: the length of stressed vowels under focus positions remains more or less unchanged. The audible appearence of lengthening may be produced by shortening other segments - in our case elements which are not directly related to focus-related stress positions.

Analyzing the data, we found out that our speaker produced these discourse-triggered differences in this way. Comparing the length of segments in non-focussed positions in detail, e.g., all segments of the subject and the finite verb, we identified more variability for the length of consonantal parts (syllable onsets - /s/, /n/, /h/) for subjects under various focus readings than for the object. In case of narrow focus on the object, some segments of the subject-constituent are obviously shortened in comparison to the two other readings[4] (cf. fig. 4 and 5).

[4] The detailed mesurements of segment duration are taken in a phonetically correct wise in both oscillogramms and spectrogramms. The former was used for the distinction of contigous

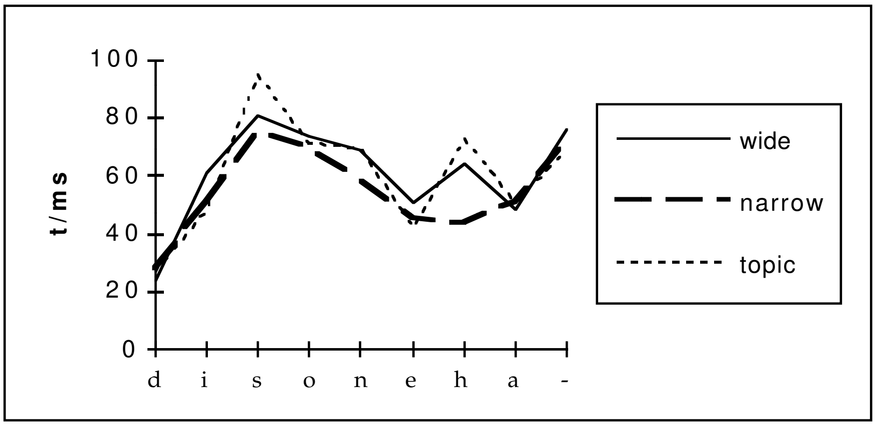

Figure 4: Segmental length of grammatical subjects

Smaller differences are observed for the segment length of the focussed constituents. Only little time variation is found for the onset of stressed syllables which are focussed. All other segments remain unaffected. As we can see in the diagram below, reduction of the absolute length for subjects in sentences with narrowly focussed objects is greater than the variation in absolute duration for the objects (cf. above).

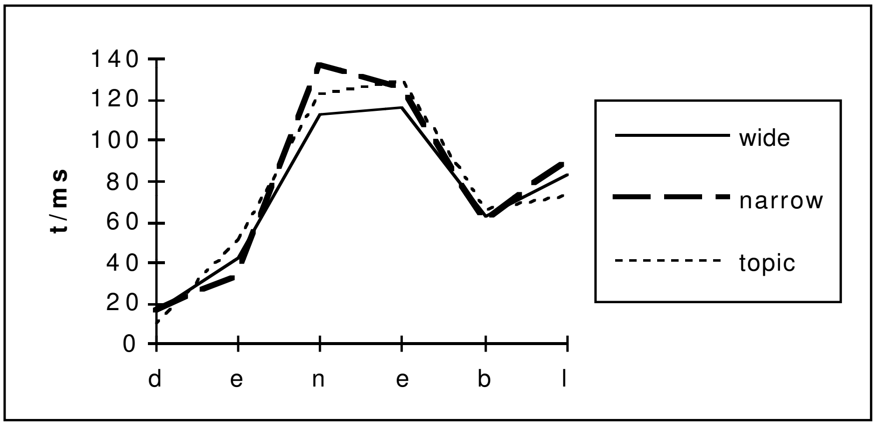

Figure 5: Segmental length of grammatical objects

Our hypothesis may be formulated in the following way: Focus positions are targets in sentence production. Our speaker of German tends to realize narrow focus as soon as possible. This goal may be reached either by variation in word order - through so-called 'topicalization' - or by shortening the segment duration of all metrically weak syllables before the phonetic target, e.g., constituents in

voiced and unvoiced segments, the latter for contigous voiced segments. Plosives are splitted, e.g.. the left peripherical /d/ contains the duration of the VOT, "-" marks the plosiv part on the right side and contains the duration between the last period of the previous vowel and the burst (the plosive aperture).

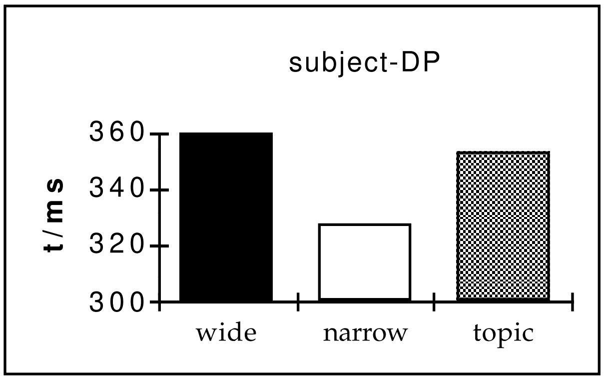

Figure 6: Constituent lengths of subjects(=DP1) and Objects(=DP2) in relation to sentence length

prefocal positions in a sentence. These observations lead us to implement this behavior of segmental length regarding the requirements of our demi-syllable synthesizer: Lexically stressed syllables are marked in the lexicon. This information is always available during sentence generation. The next step consists in weighing of stress positions in major syntactic constituents (=DPs) in the phonological component. These positions are aligned with tones according to the knowledge about discourse structure via feature interpretation: the metrically strongest are the focussed syllables. The adaptation of time patterns is realized through exploitation of already fixed stress positions.

(a) If the target, e.g., the focus accent position is not yet reached, distances between metrically strong stress positions in the prefocal part in a sentence are shortened. This is the case of wide- and narrow focus constructions.

(b) In the case of 'topicalisation', the object has 'moved' to the first position in the sentence. Thereby the focus position is reached earlier and option (a) does not apply.

3 Implementation

The *strategic generator* of the system produces sentence plans that indicate focus placement and deviations from neutral word order. These distinctions are reflected in the input to *tactical generation(Matiasek and Trost, 1996)*, as shown in fig. 7 below.

The upper box in fig. 7 shows the semantic specification common to example sentences (1)-(3). Which focus type is generated depends on the selection of the focus feature. The three lower boxes show the focus features for sentences (1)-(3) respectively.

In (3), a constraint is stated that forces the focus-projection `f-projection` to occupy the clause-initial topic-slot `topic`. If no explicit word-order constraint is provided, the grammar will fill the topic-slot by default with the subject. Specification of the `f-projection` of a clause results in the instantiation of

```
((cat s)
 (aspect perfective)
 (voice active)
 (concept vertreiben)

 (args ((actor ((concept sonne)))
        (actee ((concept nebel))))))
```

(f-type wide)	wide focus	(1)
(f-projection {↑ args actee})	narrow object-focus SVO	(2)
(f-projection {↑ args actee}) (topic {↑ f-projection})	narrow object-focus OVS	(3)

Figure 7: Input to the tactical generator

narrow focus-type `(f-type narrow)`. Alternatively, `f-type` can be specified in the input, as in (1), where wide focus is indicated (which is also the default assumed by the system when no focus specifications are provided).

To realize these semantic focus specifications, the tactical generator has to assign a nuclear pitch accent to the metrically most prominent position inside the `f-projection`. To do this, first membership of the `f-projection` is marked throughout the syntactic tree. All nodes dominated by the `f-projection` receive `(focus +)`, all others `(focus -)`. In the case of narrow focus, nuclear accent is simply assigned to the head of the `f-projection`. This is accomplished via the head-feature convention by making `(accent na)` a head feature of `f-projection`.

For wide focus, assignment of the nuclear accent follows a more complicated rule. In the case of the sentences under consideration, where a non-finite verbform remains in the clause-final verb phrase, assignment of nuclear accent can be decided by a cascaded check on properties of the clause. This procedure fixes `f-projection` in such sentences. Nuclear accent will then be realized on the head-projection of this phrase.

1. negation preceding the non-finite verb is present
   ⇒ `(f-projection neg)`
   2. the lexical verb is not a particle verb, and
      the non-finite verb is preceded by either an argument or a local adjunct
      ⇒ `f-projection` is this argument/local adjunct
      3. otherwise
         ⇒ `f-projection` is the non-finite verb

The tactical generator assigns also nuclear accents that are motivated by the melodic excursions observed in our corpus rather than by focus information. Thus, as mentioned above, in structures like (1) and (2), where the topic-slot is not occupied by a constituent carrying narrow focus, there is nevertheless a melodic excursion, which we model in tactical generationg by assigning a pre-nuclear accent to the unfocussed topic-slot:

```
(topic ((focus -)
        (head ((accent pna)))))
```

In addition, we postulated a late pitch accent on subject *die Sonne* in the OVS structure (3). This must be modeled as a second nuclear accent in the Pierrehumbert model. The assignment is done by the first two clauses of the cascaded-rule given above if `(focus -)`.

This way, grammatical processing in the tactical generator derives pitch accent types to express focus-determined prosody. In the *phonological component*, these grammatically derived pitch accents receive tonal interpretation, as High-Low (H*L) or Low-High (LH*) tones. Constituents preceding the `f-projection` are furthermore marked for shortening. The details of the phonological component of the system are discussed in (Niklfeld and Alter, 1996). The phonological component derives the input data-structures for the demi-syllable based speech-synthesizer, where the symbolic tone and timing information is expressed acoustically.

4 Conclusion

Empirical data show that linguistic categories like 'focus' influence the prosodic realization of both tonal structure and time pattern. These results of our analysis lead us to model prosody in speech synthesis by correct feature interpretation in the phonological component of our CTS. The appropriate focus feature is used for extracting relevant abstract prosodic information to guide synthesis. Better results in synthesis are expected if focus related tonal structure and time patterns are taken into account.

References

Alter, Kai. 1996. Der zusammenhang von akzentuierung und phrasierung im sprachvergleich. In Ewald Lang and Gisela Zifonun, editors, *Deutsch – typologisch*. Walter de Gruyter, Berlin, New York, pages 585–614.

Bühring, Daniel. 1995. *The 59th Street Bridge Accent: On the meaning of Topic and Focus*. Ph.D. thesis, Universität Tübingen.

Dogil, G. 1995. "alternativen" semantik und "pitch-akzent" im deutschen. ms., 17. DGfS-Jahrestagung.

Garner, Elizabeth J. 1995. The interpretation of prosodic meaning using discourse-level features for the purposes of text generation. Technical Report TR-95-38, Austrian Research Institute for Artificial Intelligence, Vienna, Austria.

Günther, C., C. Maienborn, and A. Schopp. 1994. The processing of information structure in synphonics. In P. Bosch and R. van der Sandt, editors, *Proc. of the Interdisciplinary Conference in Celebration of the 10th Ann. of the J. of Semantics "Focus & Natural Language Processing"*.

Matiasek, Johannes and Harald Trost. 1996. An hpsg-based generator for german—an experiment in the reusability of linguistic resources. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, Copenhagen.

Niklfeld, Georg and Kai Alter. 1996. Covering prosody in concept-to-speech via an extended two-level phonology component. In *Proc. SIGPHON 96*. Association for Computational Linguistics.

Pierrehumbert, Janet. 1980. *The Phonology and Phonetics of English Intonation*. Ph.D. thesis, MIT, Dept. of Linguistics.

Selkirk, E. O. 1995. Sentence prosody: Intonation, stress, and phrasing. In J. Goldsmith, editor, *Handbook of Phonological Theory*. Blackwell, London, pages 550–570.