Prosody Generation in the SmartKom Project

Antje Schweitzer & Norbert Braunschweiler & Edmilson Morais

Institute of Natural Language Processing
University of Stuttgart, Germany
schweitzer@ims.uni-stuttgart.de

Abstract

This paper describes the intonation prediction algorithm used in the speech synthesis component of the SmartKom dialog system. Both prosodic phrasing and default accent prediction operate on syntactic structure generated by a language generation module. Two configurational rules are applied to find candidate prosodic phrases. A harmonization algorithm selects the best candidates taking into account rhythm as well as a parameter specifying the optimal length of prosodic phrases. Depth of embedding and location within the syntactic structure determine the default accentuation, which can be modified depending on semantic factors.

1. Introduction

This paper describes symbolic prosody generation within the SmartKom concept-to-speech (CTS) synthesis module. We first provide some information on the project in general. In section 3, we give a detailed description of the concept input specification. Section 4 addresses the prediction of phrase breaks. Rules insert optional breaks into each utterance. A harmonization algorithm decides which of those breaks have to be discarded. Section 5 describes the default accent location procedure, deaccentuation, and the prediction of pitch accent types. We illustrate the complete intonation prediction algorithm by a short example in section 6. This is followed by a conclusion in section 7.

2. The SmartKom project

SmartKom ([6]) is a research project funded by the German Ministry of Education and Research for a period of 4 years until September 2003. There are several consortial partners involved both from industry and universities. The project is led by the German Research Center for Artificial Intelligence (DFKI). The goal of the project is to develop an intuitive multimodal dialog system which combines speech, gesture and mimics input and output. Interaction with the system is managed by a virtual communication assistant named Smartakus. It helps the user to obtain information on the cinema or TV program, control electronic devices, make reservations in restaurants, plan sightseeing tours or car routes, make phone calls, manage e-mails, etc. Smartakus also represents system output both visually and acoustically.

3. Concept input

There are two situations in which speech output is required. The first one is the interaction of Smartakus with the user. In this case, we deal with CTS: the dialog turns are generated by a language generation module and they are linguistically extensively annotated. The other situation arises when text from embedded applications has to be rendered. We will not discuss the procedure used in this case here. Details can be found in [5].

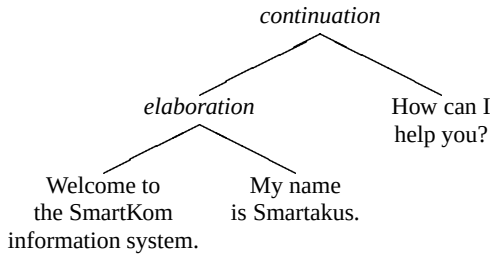


Figure 1: Discourse relations for the introductory dialog turn. The second sentence further elaborates the first one, and the third sentence is a continuation to the first two sentences.

Concepts in SmartKom contain information on various linguistic levels. Starting from the top, the highest level of annotation is discourse structure, followed by sentence level, syntactic structure and lexical level annotations. Concept structures are coded in XML.

Discourse structure. According to [4], discourse structure influences F0 register and pause duration. Discourse structure is represented by specifying discourse relations between subsequent text segments with sentences as the smallest unit. The set of discourse relations is taken from [4], although only a subset can actually be found in the short dialog turns produced by the system. An example is shown in figure 1. Discourse structure has already been specified as part of the concept structure, but it is not consistently available yet.

Sentence level. The next level of annotation is the sentence level. Sentence mode is annotated on this level. We distinguish between declarative and imperative sentences as well as between yes/no-questions and wh-questions. This kind of information is mainly required for the prediction of boundary tones.

Syntactic structure. The next lower level is syntactic structure. Syntactic trees are binary branching, and they may include traces resulting from movement of verbs or phrases. They are generated from smaller tree segments within the tree-adjoining grammar framework ([1]). An example is given in figure 2.

Additional semantic and pragmatic information is integrated into syntactic structure in the following way. For each node of the syntactic tree, its argument status can be specified. We distinguish between subjects, direct or indirect objects, prepositional objects, sentential objects, and adjuncts. Beyond that, nodes can be marked (but are not marked currently) as containing given or new information. Topic and focus can be annotated on this level as well. Pointers to other nodes can link

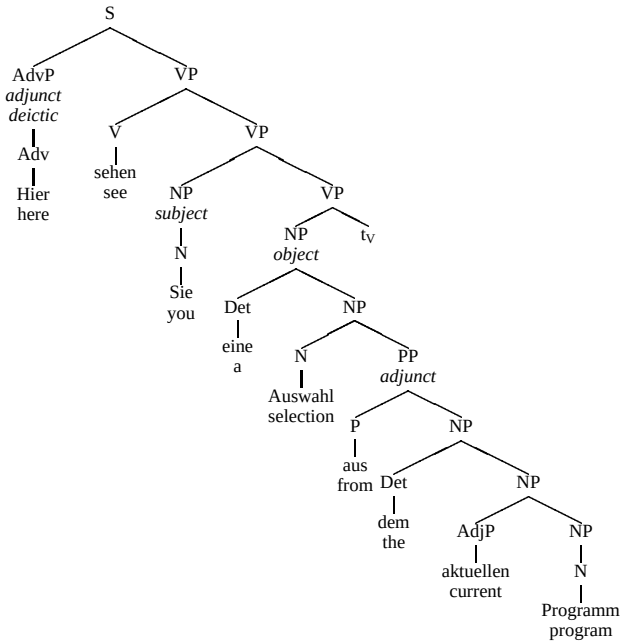


Figure 2: Integration of additional information into the syntactic structure. In this example, deixis and argument status are added. Both are indicated in italics.

contrastive elements or items in enumerations to each other. Finally, deixis is specified on the syntactic level. Deictic elements occur when Smartakus is talking about objects that are presented on the display at the same time. In this case, pointing gestures are used. The smallest units that can be marked as deictic are words, but larger constituents including complex noun phrases are also possible.

Lexical level. The lowest level of annotation is the lexical level. On this level, foreign words are marked as such. Also, for material that originates from database queries, the domain is specified. Currently, we distinguish between movie titles, actors' names, street names, etc.

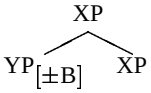
4. Prediction of prosodic phrases

The first step in prosody generation is the prediction of prosodic phrases. There are two levels of phrases: intonation phrases are terminated by major breaks and can be divided into several intermediate phrases, which in turn are terminated by minor breaks.

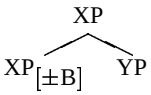
4.1. Insertion of breaks

Optional breaks are inserted by the two general rules in (1) and (2). The [±B] feature indicates that a break can be inserted at the end of the respective constituent. The two rules each have two variants, which are mirror images of each other.

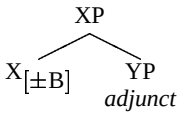
(1) a.



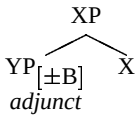
b.



(2) a.



b.



The first rule states that maximal categories that are daughters of other maximal categories can be separated from their sister node by a minor break. S constituents are treated as maximal projections. Since we use a simple syntax theory that does not distinguish XPs from XBars, rule (1) applies for any maximal projection that is not the sister of a head. Examples for the application of (1) are the insertion of boundaries between topicalized constituents and the VP as well as between adjacent complements or adjuncts within the VP.

The second rule allows breaks to be inserted between the head of a phrase and its sister node, but only if the sister node is an adjunct. Thus, phrase boundaries between a head and its argument are excluded.

Mandatory major breaks are inserted before and after S constituents. Also, deictic expressions that are accompanied by gestures are marked by preceding or following mandatory minor breaks.

The result of the phrase break insertion for the example in figure 2 is shown in (3). Mandatory major phrase breaks are at the end of the utterance, and after the deictic AdvP *hier*, indicated by the [+BB] and [+B] features, respectively. Additionally, optional phrase breaks are inserted after the NP *Sie* according to rule (1-a), and after the noun *Auswahl* according to rule (2-a). Optional breaks are marked in (3) by the feature [±B].

(3) Hier [+B] sehen Sie [±B] eine Auswahl [±B] aus dem aktuellen Programm [+BB]

4.2. Harmonization algorithm

A harmonization algorithm decides whether some breaks have to be discarded to get a smooth phrase length distribution for the complete utterance. The motivation is that we want to avoid sequences of phrases that are unbalanced in terms of number of syllables per phrase. We therefore introduce a selection step into the prediction process.

The optional phrase breaks inserted in the preceding step yield several different phrase sequences. For the selection of an optimal sequence, all possible combinations of phrases are taken into account. We first select all those candidate sequences whose mean phrase length is within an optimal range. We use a fixed optimal range of more than 4 syllables and less than 11 syllables per phrase for mean phrase length. From these candidates, we choose the one with the least variance. If no candidate is found whose mean is within the optimal range, we also accept candidates whose mean phrase length is below the optimal range. If still no candidate is found, all optional breaks are kept.

For the example in figure 2, the optimal candidate is shown in (4-a). The other candidates are given in (4-b) through (4-d). Syllable number per phrase, mean phrase length and variance are indicated in the line below each candidate. (4-b) is discarded because its mean phrase length is not within the optimal range. Of the remaining three candidates, (4-a) is chosen because it has the least variance.

(4) a. Hier [+B] sehen Sie eine Auswahl [±B] aus dem aktuellen Programm [+BB]
syllables: 1, 7, 8; mean: 5.33; variance: 9.55

b. Hier [+B] sehen Sie [±B] eine Auswahl [±B] aus dem aktuellen Programm [+BB]
syllables: 1, 3, 4, 8; mean: 4; variance: 6.5

c. Hier [+B] sehen Sie [±B] eine Auswahl aus dem aktuellen Programm [+BB]
syllables: 1, 3, 12; mean 5.33; variance: 22.89

d. Hier [+B] sehen Sie eine Auswahl aus dem aktuellen Programm [+BB]
syllables: 1, 15; mean 8; variance: 49

4.3. Boundary tones

Boundary tones are assigned for each predicted intonation phrase boundary. For sentence-internal phrase boundaries, a rising boundary tone is assigned to indicate continuation. For yes/no-questions, a rising boundary tone is used, while wh-questions are realized with a falling tone. Declarative sentences are also terminated by a falling boundary tone.

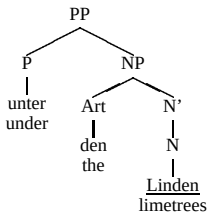
5. Pitch accent prediction

For pitch accent prediction, we first determine which element should be accented by default. Semantic factors can cause deaccentuation. Finally, we predict the accent category for each accented element.

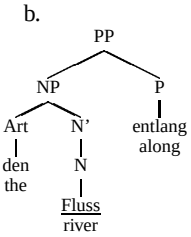
5.1. Default accent location

According to [2], the default accent is on the syntactically most deeply embedded element, as illustrated by the prepositional phrase in (5-a) and the postpositional phrase in (5-b) (from [2]). Default stress is on the underlined words.

(5) a.



b.

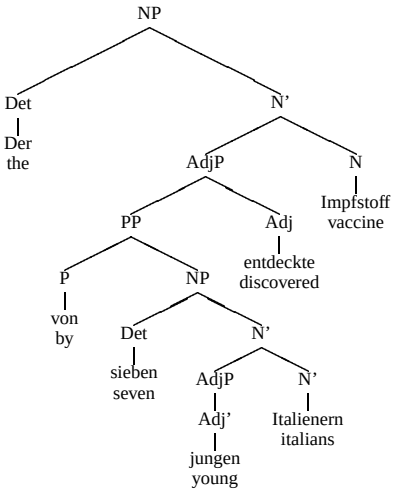


Nodes on the non-recursive side are irrelevant, as shown by (6) (from [2]): In neutral accentuation, stress falls not on the overall most deeply embedded element, *Italienern*, but on *Impfstoff*. This is because NPs are right-recursive.

Depth of embedding according to [2] is only counted on a path along the XBar axis (e.g., connecting XP and X', X' and X') and on the recursive side of each projection XP (e.g. connecting X' to an YP embedded on the left side, if XP is a left-recursive category; or connecting X' to an YP embedded on the right side, if XP is a right-recursive category). The main path of embedding is the path that reaches the top node. The overall most prominent element is the most deeply embedded element on the main path of embedding. Within constituents on the non-recursive side, depth of embedding determines the locally most prominent element within the constituent, but its depth of embedding is irrelevant for the location of the main stress.

We have to modify this procedure for two reasons. First, in the syntactic structure from the generation module, there are no XBars. This means that in our case, the main path is along an axis connecting XPs with embedded XPs, or connecting XP to a maximal projection YP on the recursive side of XP, if YP is a sister to the head X of XP. Second, large syntactic trees will usually be split up into smaller units by the phrase prediction algorithm. Within the phrases that do not contain the globally most prominent element according to the above definition, we still need to assign an accent to the locally most prominent element.

(6)



We adopted the following procedure for the prediction of accent location. All nodes within a syntactic tree are provided with a label indicating their depth of embedding. Furthermore, for each node, we count how many branches on the path from this node to the top node are neither on the XBar axis nor on the recursive side. For each phrase, the element with the smallest number of branches on the "wrong" side is accented. If there are several elements with the same number, the most deeply embedded one will be chosen. If this is again ambiguous, the last one is chosen.

If the utterance in (6) is realized as a single phrase, the noun *Impfstoff* is accented, as predicted by [2]. If it is realized in shorter phrases, e.g. with a phrase boundary between the embedded PP and the adjective, *Impfstoff* is the accented element in the second phrase. In the first phrase, the accent will be realized on the noun *Italienern*, since on the path from the top to the noun there are two branches on the non-recursive side, while there are three branches on the path to the more deeply embedded adjective.

5.2. Deaccentuation

Depending on information structure or focus-topic structure of an utterance, accentuation can deviate from the default accentuation. Although provision for the annotation of this kind of information has already been made in the definition of concepts, it is not yet available in the SmartKom system. Currently, only some specific words are defined to be inherently given. E.g., words like "now", "today", "currently" etc., and pronouns, which should necessarily be in the user's and Smartakus' common ground, are treated as given and are therefore deaccented. Currently, this is done by setting their depth of embedding to 0. Similarly, adjectives like "further", "other", etc. are interpreted as focussed. In this case, the default accent is moved to the left from the subsequent noun onto the adjective.

5.3. Accent categories

For each accented element, we predict its accent category. We use a subset of the pitch accent inventory from the German ToBI labeling system as described in [3], viz. L*H as a rising accent, H*L as a falling accent, and L*HL as an emphatic accent. The default accent category is rising. If the accent is the last accent in an intonation phrase, its category depends on the type of the following boundary. H*L accents are used before falling boundaries, and L*H accents before rising boundaries. On elements that present new information and in deictic expressions, a falling accent is used. In imperative sentences, the accented element is realized by the emphatic rise-fall L*HL.

6. An example

The complete intonation prediction algorithm is illustrated by the example in (7). An optional phrase break is inserted between the topicalized object *das Dokument* and the finite verb *wurde*. The insertion of phrase boundaries between topicalized constituents and the finite verb is very typical in German. Our algorithm selects (7-a) because (7-b)'s mean phrase length exceeds the upper limit of (less than) 11 syllables per phrase and is therefore not considered in the selection step. Otherwise, it would have won over (7-a) because its variance is smaller.

(7) Das Dokument wurde an Nils Nager verschickt.
The document was to Nils Nager sent
The document was sent to Nils Nager.

(8) a. Das Dokument [+B] wurde an Nils Nager verschickt [+BB]
syllables: 4, 8; mean: 6.00; variance: 4.00
b. Das Dokument wurde an Nils Nager verschickt [+BB]
syllables: 12; mean: 12.00; variance: 0.00

For the first phrase, the default accent is assigned to the noun *Dokument* although the path from the top of the tree to the noun contains one branch (connecting S to the NP on its left) that is neither on the recursive side nor on the XBar axis. In the second phrase, the name *Nager* is on a path exclusively along the XBar axis or along branches on the recursive side. It is therefore accented.

Since the sentence is a declarative sentence, it is terminated by a falling boundary tone. The accented element in the second phrase is assigned a falling accent for the same reason. The accent in the first phrase is predicted to be rising because the sentence continues over the intermediate phrase boundary between the two phrases.

Further examples including audio files are available at *http://www.ims.uni-stuttgart.de/projekte/smartkom/sp2002/*.

7. Conclusion

We have presented a method to predict prosodic phrases taking into account syntactic structure. First, two pairs of very simple configurational rules assign optional and non-optional phrase breaks. In a second step, a harmonization algorithm selects candidates from the set of all possible combinations of prosodic phrases in the dialog turn. Candidates whose mean phrase length lies within a given optimal range are favoured over other candidates. Out of this subset of candidates, the optimal candidate is the one with the least variance in terms of syllable length per phrase.

For accent prediction, a single rule operating on syntactic structure determines the default accent location for each prosodic phrase. Semantic factors may trigger deaccentuation. Accent types depend on the information content of the accented word, on its position within the phrase, and on sentence mode.

Future work will include the integration of additional information provided in the concept input. This concerns information structure and focus, which will be available in the concept input in the near future. Constituents larger than words will be marked as given or focussed, requiring a more elaborate mechanism for accenting and deaccenting.

Discourse structure is also expected to be available soon. Prediction of F0 registers according to [4] can then be incorporated into the prosody prediction process.

8. Acknowledgements

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the SmartKom project under Grant 01IL905K7. The responsibility for the contents lies with the authors.

9. References

[1] Becker, T., 1998. Fully lexicalized head-driven syntactic generation. In: *Proceedings of the Ninth International Workshop on Natural Language Generation*. Niagara-on-the-Lake, Ontario, Canada.

[2] Cinque, G., 1993. A null theory of phrase and compound stress. In: *Linguistic Inquiry*.

[3] Mayer, J., 1995, Transcribing German Intonation - The Stuttgart System. Technical Report, University of Stuttgart.

[4] Möhler, G.; Mayer, J., 2001. A discourse model for pitch-range control. In: *Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis*. Perthshire, Scotland.

[5] Schweitzer, A., 1999. Intonationsbestimmung auf der Basis von Wortklassen. Master thesis, University of Stuttgart.

[6] The Smartkom Project, *http://www.smartkom.org*.