Prosody Generation in Text-to-Speech Conversion Using Dependency Graphs

Anders Lindström *Ivan Bretan* *Mats Ljungqvist*

Telia Research AB, S-136 80 HANINGE, SWEDEN

ABSTRACT

The present paper addresses the problem of prosody assignment in the context of a system for text-to-speech conversion of Swedish. The question of what type of textual analysis is needed for prosodic purposes is discussed and the use of Dependency Graphs to this end is proposed. It is argued that this type of analysis provides sufficient syntactic information so that prosodic phrasing and prominence assignment algorithms can be successfully applied. The ability of the system to generate acceptable prosodic descriptions is evaluated by comparison with examples of prosodically analysed naturally spoken sentences.

1. INTRODUCTION

Speech synthesis technology has progressed remarkably over the past few years, especially with regard to segmental naturalness. As sound quality is improved, however, the lack of proper prosodic phrasing and prominence assignment becomes all the more obvious. Most commercial text-to-speech systems today employ rather simplistic methods, typically assigning a default sentence accent based on the content/function word distinction. The present paper addresses the problem of prosody assignment in the context of a system for text-to-speech conversion of Swedish. The goal is to produce an automatic analysis of arbitrary (even ill-formed) text which can be converted into a symbolic prosodic representation of the utterance, in terms of phrasing and accentuation.

The question of precisely what type of analysis is required for this is not without controversy. Not long ago, it was generally believed that prosodic performance could be predicted from syntactic structure alone. However, results from the field of psycholinguistics, such as those presented by Gee & Grosjean (G&G) [6] indicated the existence of so-called prosodic *performance structures* that were not directly derivable from syntax. Rather, it must be acknowledged that semantics and discourse structure are among the most important factors affecting intonation and phrasing. Using an earlier version of our TTS system, it was shown how such factors can, to some extent, be taken into account in the case of restricted texts [8] and possibly even in so-called unrestricted texts [11]. Our working assumption is that factors relating primarily to semantics, pragmatics and the functional roles of the utterance are manifested indirectly in the speech signal in terms of F_0 movements, segment durations, pausing, pre-boundary lengthening, segmental reductions, coarticulation, etc.

Within this class of algorithms there seems to have been a recent, general trend to separate syntactic knowledge from, for example, rules that take into account rhythmic patterns, semantic focus etc. Systems with such clearly separated syntactic components have been proposed by e.g. Black & Taylor [2] and Dirksen [5]. This separation is laudable, as the mapping between syntactic structures and phonological structures, although not straight-forward, is clearly needed, as noted by many, e.g. G&G.

One remaining problem is to find a grammatical representation which provides sufficient information for prosodic phrasing and prominence assignment algorithms to be successfully applied both in the absence and presence of discourse-related information. In this paper, we propose Dependency Grammar to be well suited for this task, and demonstrate this by comparing dependency analyses of example sentences with intonational and temporal representations of read speech material in Swedish.

2. PARSING FOR PROSODY USING DEPENDENCY GRAPHS

We propose a dependency-based syntactic representation level which seems more appropriate for phonological phrase formation than traditional phrase-structure trees. In fact, when examining the requirements on syntactic representations placed by the ϕ algorithm—presented by G&G and put to practical use by Bachenko & Fitzpatrick (B&F) [1]—not only are syntactic *phrases* assumed to be identified; in addition, the notion of *head* is central. The head is the "core" of the phrase, the word around which the other words of the phrase are organised.

The importance of the head–modifier distinction in prosodic generation is reflected by the fact that B&F define phonological phrases as consisting of a syntactic head and the material that intervenes between it and a preceding head: "Every phonological phrase boundary thus marks a syntactic head as well as the site of a possible prosodic boundary in speech."

In Fig. 1, phrase boundaries that B&F would posit are shown for one of the example utterances that we have been examining. Note

Hela	*gruppen*	*gav upp*	*ett*	*beundrande*	*tjut*	*inför*	*den*	*enastående*	*prestationen.*
Adj	N-def	V-past	Det	V-pc	N	PP	Det	Adj	N-def
The entire	group	emitted	an	admiring	cry	when presented with	the	amazing	feat.

Figure 1: Phrase boundaries according to Bachenko and Fitzpatrick [1].

that the particle *upp* is bundled with the verb and the preposition *inför*, although a head, goes with its modifier NP according to the principles of phonological word formation assumed by B&F.

In the original work of G&G, a fair amount of syntactic processing was taking place in the construction of their ϕ-phrases, but in a largely procedural manner. Just like Dirksen explicitly states [5], we instead want to introduce declarativeness into the syntactic part of phonological phrase boundary recognition.

Taking as the point of departure the need for a declarative framework for syntactic analysis where the head–modifier relation is central, we have arrived at dependency grammar [9]. Syntactic analyses representing dependency differ from ones based on constituency by being oriented towards the *dependent-on* relation between the so called regent and its dependents (i.e., a part–part relation), as opposed to the traditional *consists-of* relation of phrase structure trees (a part–whole relation). The regent of the phrase is basically the head, and the dependents are its adjuncts and complements. Since the notion of head is crucial when assigning prosodic phrase boundaries, dependency analyses are more useful than conventional constituency analyses. Every node in a dependency tree consists of a head and a number of dependents, and can thus be taken as a candidate for a prosodic phrase.

More specifically, we propose a dependency-based unificational grammar formalism loosely based on the formalism described by Steimann and Brzoska [13]. Using a robust chart-style parser, packed dependency graphs representing all possible analyses are built from rules in this formalism. The dependency grammar rules, which have a declarative reading, have the following format:

```
Regent(HF1=HFValue1,...,HFN=HFValueN) :>
   D1(D1F1=D1FValue1,...,D1FN=D1FValueN),
   ...
   DM(DMF1=DMFValue1,...,DMN=DMFValueN),
   {OrdRel1,...,OrdRelN}.
```

where `HF` = Head Feature, `DMF` = Feature for Dependent M, and `OrdRelN` = Ordering Relation N.

Disjunctions of dependents can also be specified as well as feature negation. The optional ordering relations specify the internal ordering constraints among the dependents and regent. A sequence of tagged lexical items constitutes the input to the dependency parser. The tagger used was the Xerox part-of-speech tagger XPOST [3], modified for Swedish [11].

Fig. 2 shows a typical dependency analysis of the sentence in Fig. 1. Identifying candidate phonological phrase boundaries using this representation as input is straight-forward: they are found following each regent word (with the exception of the preposition *inför* which is treated specially in accordance with G&G).

The analysis in Fig. 2 actually gives us more structure than would be required to construct the phonological phrases mentioned above. Furthermore, this level of analysis is difficult to obtain without proper mechanisms for disambiguating attachments (in the example, attaching the PP *inför den enastående prestationen*). Interestingly enough, this level of attachment is *not* needed to build the phonological phrases of B&F. We can take advantage of this fact when using the rule processor we have developed to support the dependency grammar format.

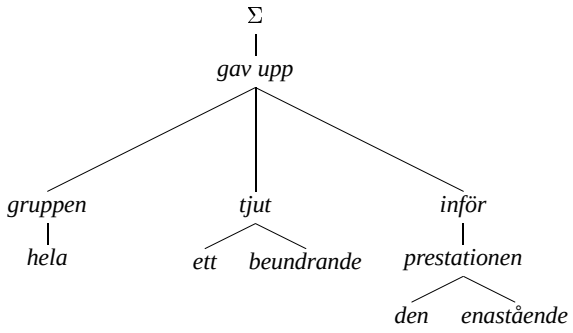

Figure 2: Dependency analysis.

The first stage of parsing results in a packed dependency graph where all possible analyses are represented. The second processing stage therefore deals with unpacking the graph. This can be done in a number of ways, for example by applying probabilistic collocation criteria. The unpacking procedure most suitable for our purposes does away with all ambiguities, thus producing a "safer", but flatter analysis. For example, in such an analysis an ambiguously attaching PP would be considered independent from the heads it could possibly modify. Of course, if this flatter analysis always is the one desired, it is a waste of time to go through the costly process of constructing all the different analyses representing the different attachment possibilities. Instead, the grammar itself can automatically or manually be transformed into an efficient "flat" grammar which will not generate any structural ambiguities. Due to the robust bottom-up character of the parser, an analysis is always produced, even when an unambiguous connected dependency graph spanning all the words of the utterance can not be obtained. Using a flat grammar, the analysis of the sentence in Fig. 1 would look as in Fig. 3.

Dependency grammars are weakly equivalent to phrase structure grammars, and although an elegant way of respecting the importance of the head–modifier distinction in syntax, we might as well have made use of a constituency-based grammar enhanced to take into account the notion of head (such as HPSG). However, there are other characteristics of dependency representations that make them particularly interesting for prosodic generation, most notably their close mirroring of grammatical functions. Merle Horne [7] claims

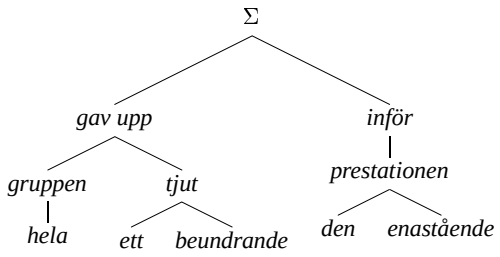

Figure 3: Flattened dependency analysis.

```
[Sigma [[exklusivt [så [inte]]]
       [kanske]
       [eller [märkvärdigt [speciellt]]]
       [men [gott]]
       [gott [mycket]]]
```

Figure 4: Dependency analysis in bracket notation of the Swedish sentence "Inte så exklusivt kanske eller speciellt märkvärdigt, men gott, mycket gott.".

that the position of sentence accent is conditioned by a hierarchy of grammatical functions (together with the notion of new and given information):

predicate complement $>$ subject $>$ predicate

Thus, disregarding previously given information, by default sentence accent would be assigned to the predicate complement (typically an object), or, if no complement exists, the subject. Finally, in a sentence with only a predicate, this would be accented. Yet another benefit of the dependency approach becomes evident if it is necessary to project the sentence accent onto the heads of the dependents of the accented regent, as Horne predicts.

3. COMPARISON WITH READ SPEECH

In order to examine the feasibility of a dependency-graph based approach to prosodic grouping, comparisons were made with natural speech samples.

Acoustic cues that are known to signal prosodic phrase boundaries include durational ones such as preboundary lengthening, pausing and speaking rate changes as well as tonal ones such as fundamental frequency resets. All these cues contribute to the perceived boundary strength. Generally, the more cues that are present, the stronger the perceived boundary. There also seems to be a high correlation between perceived boundary strengths in normal speech and in delexicalized speech [4], indicating that subjects can successfully distinguish different boundary strengths even without access to the semantic content.

In this work we have adopted "normalized duration" [15] as a way of detecting prosodic phrase boundaries by measuring preboundary lengthening effects. Normalized duration ($\tilde{d}$) is a measure of the deviation from the mean duration of the same phoneme, expressed in number of standard deviations. The phoneme statistics are gathered on a per speaker basis[1] and compensated for speaking rate changes between sentences. The segmentation used for obtaining the phoneme durations was generated automatically using the SRI Decipher speech recognizer [14] trained for Swedish.

Figures 5 and 6 show normalized durations ($\tilde{d}$) averaged using a

[1] Mean durations are computed over all instances of a phoneme, including lengthened ones. Mean durations will therefore be larger than the duration of unlengthened segments [15]. It would be more appropriate to calculate the statistics only on the unlengthened segments.

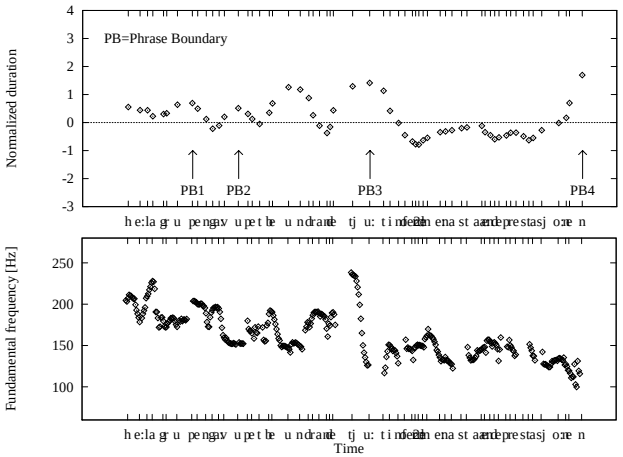

Figure 5: Normalized duration and F_0 plots of the Swedish sentence "hela gruppen gav upp ett beundrande tjut inför den enastående prestationen" ("The entire group emitted an admiring cry when presented with the amazing feat")

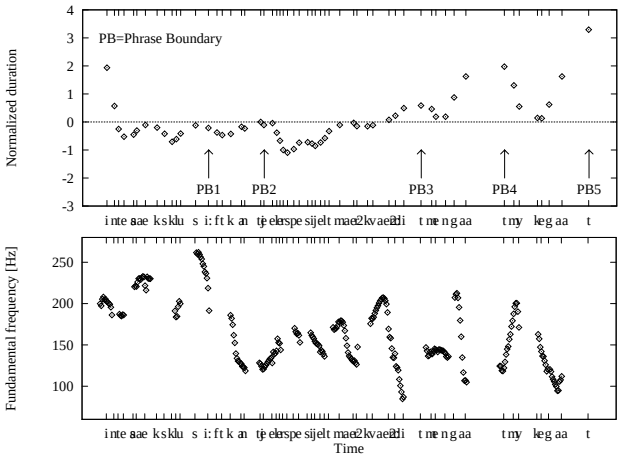

Figure 6: Normalized duration and F_0 plots of the Swedish sentence "inte så exklusivt kanske eller speciellt märkvärdigt, men gott, mycket gott" ("Not so exclusive perhaps or anything special, but delicious, very delicious").

3-point window, and fundamental frequency contours (F_0) for the sentences analyzed in Figs. 3 and 4 spoken by a professional female speaker (an actress). The locations of possible prosodic phrase boundaries predictable from the dependency graphs 3 and 4 are indicated with arrows in the figures. In most cases these locations correspond well to local maxima in the $\tilde{d}$-plots. The $\tilde{d}$-plot in Fig. 5 does, however, exhibit a major peak associated with the word "beundrande" ("admiring"). This peak is not accompanied by the typical F_0 gesture normally associated with sentence accent, which is instead realized at the following word ("tjut"). This realization, combined with the durational cues of "tjut", seem to mark a phrase boundary, while the durational gesture of "beundrande" can be interpreted as corresponding to an emotionally related emphasis rather than to a prosodic phrase boundary.

It is also interesting to note, in the utterance in Fig. 6, the relatively high speaking rate of the introductory, parenthetical material prior to PB4, the semantic focus of the sentence: "gott" ("delicious").

4. CONCLUDING REMARKS

The scheme described in this paper features a prosodically motivated analysis, intended for use in text-to-speech conversion: the output of a morphosyntactic component (a part-of-speech tagger) is used to build dependency graphs, i.e. structures that reflect the head-modifier relations among constituents. A dependency-oriented grammar is used to identify the head-modifier relationships in the output from the morphosyntactic component, thereby specifying the structure of dependency graphs while resolving certain surface ambiguities. The resulting complete graph or remaining maximal subgraphs constitute the working material for prosodic generation, possibly paired with available information regarding the current discourse. The feasability of this scheme was demonstrated by comparison with examples of read Swedish sentences, that were analysed with respect to intonational and durational characteristics.

The Dependency Analysis task could be even further simplified by using a tagging scheme such as that of Constraint Grammar [10], since that tagging scheme yields partial information about the dependencies among the words.

The resulting structure-generating component is meant to be integrated as the last stage of the text analysis component of a text-to-speech system. A subsequent component assigns prosodic boundaries and degree of prominence based on the dependency analysis. The resulting structure is then passed on to a component, where a set of rules is used to turn this symbolic prosodic notation into temporal and intonational realizations.

It is not quite clear whether final lengthening really is an independent marker for the end of a phrase. It has been suggested [12] that this phenomenon may instead be a consequence of the focus position and the associated F_0 change. Our findings are compatible with both views since the sentences chosen for the present study do not contain any out-of-focus phrase boundaries. However, the coupling between durations and F_0 gestures needs to be more thoroughly investigated.

5. REFERENCES

1. J. Bachenko and E. Fitzpatrick. A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 16:155–167, Sept. 1990.
2. A. W. Black and P. Taylor. Assigning intonation elements and prosodic phrasing for english speech synthesis from high level linguistic input. In *Proc. of ICSLP*, pages 715–718, Yokohama, 1994.
3. D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun. A practical part-of-speech tagger. In *Proc. of the 3 rd Conference on Applied Natural Language Processing*, pages 133–140, Trento, Italy, 1992.
4. J. R. de Pijper and A. A. Sanderman. On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *J. of the Acoustical Society of America*, 96(4):2037–2047, Oct. 1994.
5. A. Dirksen. Accenting and deaccenting: a declarative approach. In *Proc. of the 14 th COLING*, Nantes, France, 1992.
6. J. P. Gee and F. Grosjean. Performance structures: A psycholinguic and linguistic appraisal. *Cognitive Psychology*, (15):411–458, 1983.
7. M. Horne. *Towards a Discourse-Based Model of English Sentence Intonation*. PhD thesis, Lund University, 1987.
8. M. Horne, M. Filipsson, M. Ljungqvist, and A. Lindström. Computational modelling of contextual coreference: Implications for Swedish text-to-speech. In P. Bosch and R. van der Sandt, editors, *Proc. of Focus & Natural Language Processing*, volume 1. Intonation and Syntax, 1994.
9. R. A. Hudson. Constituency and dependency. *Linguistics*, 18:179–198, 1980.
10. F. Karlsson, A. Voutilainen, J. Heikkilä, and A. Anttila, editors. *Constraint Grammar: a language-independent system for parsing unrestricted text*. Mouton de Gruyter, Berlin—New York, 1995.
11. A. Lindström, M. Horne, T. Svensson, M. Ljungqvist, and M. Filipsson. Generating prosodic structure for restricted and "unrestricted" texts. In *Proc. of the 13th Intl. Congress of Phonetic Sciences*, Stockholm, 1995.
12. B. Lyberg and B. Ekholm. The final lengthening phenomenon in Swedish—a consequence of default sentence accent? In *Proc. of the 3rd Intl. Conf. on Spoken Language Processing*, pages 135–138, Yokohama, 1994.
13. F. Steimann and C. Brzoska. Dependency unification grammar for Prolog. *Computational Linguistics*, 21(1):95–102, 1995.
14. M. Weintraub, H. Murveit, M. Cohen, P. Price, J. Bernstein, G. Baldwin, and D. Bell. Linguistic constraints in hidden Markov model based speech recognition. In *Proc. of ICASSP*, pages 699–702, Glasgow, Scotland, 1989. IEEE.
15. C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price. Segmental durations in the vicinity of prosodic phrase boundaries. *J. of the Acoustical Society of America*, 91(3):1707–1717, Mar. 1992.