Computer Speech and Language (1992) **6**, 77–98

The derivation of prosody for text-to-speech from prosodic sentence structure

Hugo Quené and René Kager

Research Institute for Language and Speech, Rijksuniversiteit Utrecht, Trans 10, NL-3512, JK Utrecht, The Netherlands

Abstract

Suprasegmental phenomena in synthetic speech should reflect the linguistic structure of the input text. An algorithm is described, which establishes the prosodic sentence structure (PSS). This can be achieved without exhaustive syntactic parsing, using a dictionary of 550 function words. Subsequently, phrase and accent locations are derived from the PPS; accentuation is also affected by some semantic and contextual information. Comparison of the resulting sentence prosody with that of a human (professional) speaker shows that more detailed syntactic analysis may be necessary. Most of the accentuation errors are caused by semantic, pragmatic and contextual factors. These factors can only partly be imitated (using heuristics), since the relations between linguistic representations and real-world knowledge are not yet fully understood.

1. Introduction

There is ample evidence that adequate suprasegmental cues make a sentence easier to perceive and to comprehend (e.g. Collier & 't Hart, 1975; Wingfield, 1975; Nooteboom, Brokx & De Rooij, 1978; Cutler, 1982; Nooteboom, 1985; Cutler & Clifton, 1984; Nooteboom & Kruyt, 1987). Using suprasegmental cues, the listener can extract the linguistic structure of the message. *Pauses* between (linguistically) coherent word groups, for example, help the listener in three ways (Scharpff & Van Heuven, 1988). First, word segmentation is facilitated, because pauses coincide with word boundaries; the positive marking of these word boundaries in the speech stream reduces the ambiguity with regard to word segmentation. Second, the continuous speech signal is divided into coherent word groups or chunks of acoustic–phonetic information, which increases the intelligibility of the speech signal. Third, the listener is provided with some extra processing time between these coherent word groups. Likewise, *accentuation* guides a listener's attention to the words which are considered important by the speaker, and which are acoustically most reliable.

These perceptual functions of sentence prosody become even more important when the speech signal is less redundant, providing fewer segmental cues to the intended speech sounds, words and meaning. Obviously, artificial speech (such as the output of a text-to-speech conversion system) lacks the normal degree of acoustic–phonetic redun-

0885-2308/92/010077+22 $03.00/0

dancy, because it is not known which (and how) phonetic details should be implemented. Consequently, correct suprasegmental cues may significantly improve the perception and comprehension of synthetic speech. Text-to-speech systems should aim at producing a natural sentence prosody, in order to compensate for their reduced speech quality.

In the production of natural speech, several suprasegmental phenomena depend on the intended *phrasing* of a speech utterance: a linguistic boundary may be marked by means of an appropriate F_0 movement (Collier & 't Hart, 1975; Cooper & Sorensen, 1977), a silent interval (Goldman-Eisler, 1972), lengthening of the preceding speech sounds (Klatt, 1975, 1976), blocking of coarticulation and sandhi (Cooper & Paccia-Cooper, 1980), etc. These phenomena can be seen as *phonetic* correlates of the *abstract* notion "phrase boundary". Together, they indicate the intended division of the speech utterance into phrases.

Similarly, several suprasegmental phenomena depend on the *accent* of a word, i.e. on its relative prominence. A speaker can convey this word prominence by means of appropriate F_0 movements (Collier & 't Hart, 1975), in combination with higher intensity (Lehiste, 1970), longer duration (Klatt, 1975, 1976) and less vowel reduction (of the stressed syllable within the accented word; Koopmans-Van Beinum, 1980). Again, these phenomena can be seen as phonetic correlates of the more abstract notion "accent" (Nooteboom & Kruyt, 1987).

Thus, various phonetic aspects of sentence prosody depend (to a great extent) on the abstract linguistic notions *phrasing* and *accent*. From these latter two, a text-to-speech system can (in theory) derive many suprasegmental phenomena in the output speech: segmental durations, location and duration of silent intervals, F_0 movements (e.g. Van Wijk & Kempen, 1985), vowel reduction, intensity pattern, sandhi and coarticulation, etc. Consequently, a high quality text-to-speech system should attempt to establish both of these abstract prosodic phenomena, and to convert these into correct suprasegmental phenomena.

In natural speech, the (abstract) phrasing and accentuation are assumed to be related to the linguistic structure of an utterance. According to non-linear phonological theory, abstract sentence prosody does not depend directly on the syntactic surface structure, but rather on the related *prosodic sentence structure* (Nespor & Vogel, 1982, 1986; Gee & Grosjean, 1983; Selkirk, 1984, 1986). In addition, abstract accentuation is also affected by: (a) the *thematic* structure, which specifies the semantic constituents functioning as predicate, argument and modifier (Gussenhoven, 1984); as well as by (b) the *focus* structure, which relates the syntactic constituents to be emphasized with specific accents (Baart, 1987).

Extending this line of thought, we assume that accentuation and phrasing can both be derived from the prosodic sentence structure, provided that thematic structure and focus information are taken into account. To illustrate matters, our view of the various levels of sentence prosody is represented in Fig. 1.

The aim of the present paper is to investigate whether this abstract prosody can be derived automatically for text-to-speech conversion. To this end, we have developed an experimental algorithm (PROS). This algorithm attempts to derive both phrasing and accentuation from the prosodic sentence structure, in combination with some approximation of focus information and thematic sentence structure. First, a hybrid prosodic sentence structure is established (hence PSS). Although the resulting structure comes close to the prosodic sentence structure proposed by Nespor and Vogel (1982, 1986; explained below in more detail), thematic relations are also taken into account, as far as

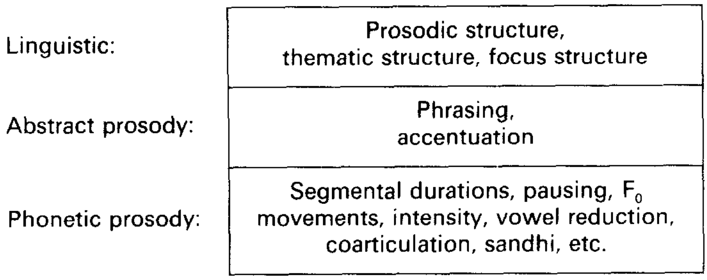
Linguistic:
Prosodic structure, thematic structure, focus structure
Abstract prosody:
Phrasing, accentuation
Phonetic prosody:
Segmental durations, pausing, F_0 movements, intensity, vowel reduction, coarticulation, sandhi, etc.

Figure 1. Three levels of sentence prosody.

these can be approximated: separate thematic constituents (viz. predicates, arguments, modifiers) usually correspond to separate prosodic domains. Subsequently, accentuation and phrasing are derived by means of this sentence structure (as well as by means of additional information) and inserted as abstract prosodic markers into the text sentence. In short, the algorithm derives abstract prosody from a prosodic sentence structure.

This algorithm is implemented in a Dutch text-to-speech system (Te Lindert, Doedens & Van Leeuwen, 1989). Other components convert its output (viz. abstract prosody) to adequate phonetic suprasegmentals: on the basis of the resulting phrase boundaries and accent locations, silence segments are inserted in the phoneme string, durations of segments (i.e. diphone fragments) are adjusted, and an F_0 contour is calculated (more details regarding the latter are given by 't Hart & Collier, 1975; Terken & Collier, 1989).

The algorithm presented here is the latest member of a steadily growing family of similar algorithms (Kulas & Rühl, 1985; Allen, Hunnicutt & Klatt, 1987; Ladd, 1987; Monaghan, 1990*a*,*b*,*c*; Quazza, Varese & Vivalda, 1989; O'Shaughnessy, 1989; Carlson, Granström & Hunnicutt, 1989; Bailly, 1989; Hirschberg, 1990; Bachenko & Fitzpatrick, 1990). The main feature shared by these algorithms is the acknowledgement that suprasegmental phenomena cannot be derived from syntactical information alone; non-syntactical information also plays an important role in determining sentence prosody (e.g. phrase length, reference to "given" vs. "new" concepts). Since syntactic factors can be "overruled" by other types of factors, an exhaustive syntactic analysis is generally superfluous. In spite of these similarities, however, our method differs in several respects from those mentioned above (apart from being developed for Dutch). These differences mostly stem from our distinction between three separate tasks in generating sentence prosody, viz.: (a) derivation of sentence structure; (b) derivation of abstract phrasing and accentuation; (c) generation of suprasegmental parameters.

This three-step approach offers many advantages over other methods, where two or more of these tasks are combined. First, it seems to be a better approximation of human speech production. The suprasegmental phenomena associated with phrasing and accentuation (pausing, intonation, segment duration, etc.) are obviously intertwined. Hence, both phrasing and accentuation are likely to be derived from the same sentence representation. This representation should combine syntactic, semantic and rhythmic relations; the adjusted PSS matches these requirements—as opposed to syntactic surface structure, tone group structure or phrase structure. Second, the PSS may also be used to control other suprasegmental phenomena in the resulting speech signal, like those indicated in Fig. 1. Third, there will be no conflicting suprasegmental cues in the resulting synthetic speech: all cues indicate the same abstract prosody and linguistic structure. This makes the sentence structure relatively easy to retrieve.

2. Prosodic structure: theory

From many languages, it is known that sandhi rules (i.e. rules of phonological adjustment between words) have their own specific domains of application. These domains are not necessarily isomorphous to syntactic constituents. Among others, Nespor and Vogel (1982, 1986) and Selkirk (1984, 1986) have made proposals as to the mapping between syntactic constituents and *prosodic domains*. Two prosodic domains attested in this sense are the phonological phrase (Phi) and the intonational phrase (Int). These domains are part of a hierarchical tree structure, viz. the prosodic sentence structure (PSS).

In languages such as English and Dutch, the *Phi* domain (or phonological phrase) is built around a lexical head, i.e. a content word (hence CW, i.e. either noun, verb, adjective or adverb) which is the head of a syntactic constituent. It includes left-hand specifiers of the lexical head, as well as all left-hand function words (hence FWs, i.e. prepositions, conjunctions, complementizers, copula, etc). The (final) lexical head of each Phi is the *prosodic head*; this word plays an important role in accentuation (see Section 3.3 below).

The next higher prosodic constituent is the *Int* domain (or intonational phrase). This constituent is constructed by grouping adjacent Phi domains. Hence, a whole Phi domain is always contained within a single Int domain. In addition, however, important syntactic breaks are also respected. In general, each syntactic constituent which is attached to any S-node (in the syntactic surface structure) establishes a separate Int domain. Consequently: (1) displaced syntactic constituents; (2) (most) subordinate clauses; and (3) parentheticals, are all separate Int domains.

In the following example, the Phi and Int domains are illustrated in a flat representation (where '##' indicates an Int-boundary, and '#' a Phi-boundary). These examples clearly demonstrate that prosodic domains do not necessarily correspond to syntactic constituents.

## Kasyapa's great war elephant # turned aside (1)
to avoid # a patch # of marshy ground

## de computer # spreekt # tot de bemanning (2)
## op de betweterige # en begrijpende toon
## die we kennen # uit de zachte sector ##
"## the computer # speaks # to the crew
## in the pedantic # and understanding tone
## which we know # from the soft sector [of social workers] ##"

Prosodic domains tend to be of equal length, and their length increases in faster speech. To account for these effects, separate rules restructure the prosodic domains. An optional rule joins a Phi consisting of only the lexical head to the Phi to its left, under some syntactic conditions. Very short Ints can be eliminated by merging them with adjacent Ints, and very long Ints are broken down into shorter ones.

3. Automatic prosodic analysis

3.1. Introduction

According to the linguistic theory described above, the prosodic sentence structure is derived from the syntactic (surface) sentence structure. Hence, a syntactic parsing is necessary for any text-to-speech system performing prosodic analysis.

However, this linguistically motivated method cannot be applied to *automatic* prosodic sentence analysis. First, there is no parser available which performs satisfactorily for our purposes. Such a parser must be able to analyse any text, at a speed which exceeds the average speaking rate. This task requires a large set of syntactic rules, as well a large lexicon. At this moment, such a system is not (yet) available for Dutch. Second, if such a parser did exist, it would run into great difficulties when analysing syntactically ambiguous sentences like the following:

I (have mown) (the lawn with the flowers). (3a)
*I (have mown) (the lawn) (with the flowers). (3b)

het was (ondanks de luchtverversing) (door de tv-lampen snikheet). (4a)
"it was in-spite-of the air-conditioning because-of the TV-lights suffocatingly-hot".
*het was (ondanks de luchtverversing door de tv-lampen) (snikheet) (4b)
"it was in-spite-of the air-conditioning by-means-of the TV-lights suffocatingly-hot".

The constituents are identical in these two analyses but the inter-constituent syntactic and thematic relations differ. The different thematic relations result in different accentuations, as will be explained in the Theory subsection of Section 3.3.2. Solving this type of thematic ambiguity requires a semantic and pragmatic analysis; the parser must "know" that one cannot use flowers to mow a lawn, and that TV lights produce heat rather than fresh air. Again, no system exists for this type of sentence analysis.

For these reasons, many researchers have avoided an exhaustive syntactic parsing of the input sentence. Quite often, this drastic step results in obviously incorrect suprasegmental cues (Klatt, 1987). In the present approach, however, suprasegmental cues are *not* derived directly from the (incomplete) syntactic structure, but instead from the PSS (via abstract prosody). We assume (on theoretical grounds) that prosodic phenomena can be derived correctly from the PSS. This implies that a partial syntactic analysis need not deteriorate the adequacy of suprasegmental cues, as long as a correct PSS underlies the latter. This prosodic structure can be approximated from an incomplete syntactic structure. Rather than deriving suprasegmental cues directly from a partial syntactic analysis, the latter is better aimed at establishing a correct PSS, from which suprasegmental cues can be derived.

It should be noted, however, that the resulting PSS can only *approximate* the theoretical prosodic structure, since not all relevant syntactic information is available for the prosodic analysis. Some of this syntactic information is of vital importance. For example, the main verb (or verb group) in a sentence must be identified (cf. O'Shaughnessy, 1989). This word (group) establishes a predicate constituent, corresponding to a separate Phi domain (Gee & Grosjean, 1983). This Phi domain may separate the subject and object arguments of the predicate. Likewise, subordinate sentences must be identified, because they usually establish separate Int domains (see Section 2). In Section 3.2.1 below we will argue that the information required for prosodic domain construction (PSS) can often be derived from the syntactic word class. The present approach is illustrated in Fig. 2.

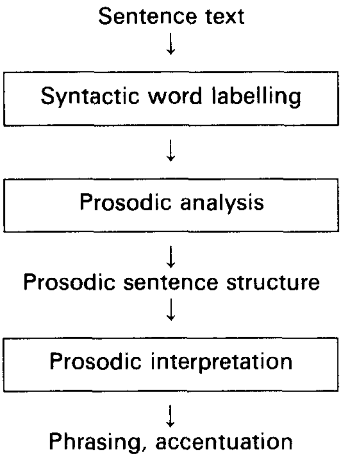

Figure 2. Alternative method for deriving the prosodic sentence structure (PSS) directly from an input sentence.

As a first step, then, the words constituting the sentence must be provided with a syntactic label. Subsequently, the PSS is derived from both the orthographic input sentence (text string), and from the syntactic labelling of its constituent words. Finally, (abstract) phrase boundaries and accents are derived from the PSS, as well as from the syntactic word labelling. (A full listing of the rules, in quasi-SPE format, is available from the first author.)

3.2. From text to PSS

3.2.1. Introduction

Two types of syntactic information are indispensable for prosodic sentence analysis. First, lexical heads must be identified. This is achieved by identifying each word as either CW or FW. In theory, each last CW preceding an FW (or sentence boundary) constitutes a lexical/prosodic head (carrying abstract accent; cf. Kulas & Rühl, 1985). The CW-or-FW status of a word is determined by means of lexical look-up, in an "FW dictionary", which contains all Dutch FWs (554 types). In addition, the dictionary contains another 300 CWs which behave anomalously for one reason or another. For all words, the syntactic word class is also specified. This syntactic labelling serves two purposes: (1) determining the syntactic labels of the words not found in the dictionary; and (2) the construction of prosodic domains, as explained below. (More details on this dictionary are given in the following section.)

Second, syntactic phrasing must be known, since it may be relevant for prosodic domains. As explained before, however, an exhaustive syntactic analysis appears to be superfluous for establishing prosodic domains: it is not necessary to determine the structural relations between the words in a sentence. Returning to (1), for example, it is not necessary to decide which is the internal structure of the subject NP:

(Kasyapa's (great (war elephant))). (1a)
(Kasyapa's ((great war) elephant)). (1b)
((Kasyapa's (great war)) elephant). (1c)

Instead, it suffices to determine that these four CWs together establish a single constituent, which in turn establishes a separate Phi domain. This information can be derived from the syntactic class of the words in (1):

Kasyapa's great war elephant # turned aside . . .
N A N N # V A . . . (1)

The noun **elephant** and the inflected verb **turned** cannot belong to the same syntactic constituent, as the noun cannot be a specifier to an inflected verb. Hence, **elephant** must be lexical head of the constituent preceding the verb **turned**; i.e. the two words must belong to separate syntactic constituents. Since both words are lexical heads of syntactic constituents, they cannot both belong to the same Phi domain. Consequently, a Phi boundary must separate these two words.

In more general terms, our approach is based on the fact that within a syntactic constituent, some possible sequences of syntactic labels are allowed, while others are not. If constraints on the possible label sequences are violated, then we may assume a syntactic boundary. Hence, syntactic constituency can be derived from the sequence of syntactic word labels (O'Shaughnessy, 1989). Not all syntactic boundaries, however, are equally important for the PSS. For the purposes of sentence prosody, it generally suffices to demarcate those syntactic constituents which are respected by prosodic domains, while ignoring other syntactic structures.

Because the algorithm demarcates prosodic domains on the basis of both the CF/FW distinction and syntactic word labels, it can deal with a large variety of input text types. For its development, a large corpus of newspaper text was used (approximately 450 000 word tokens). This corpus consists of editorial news bulletins on numerous subjects, as published in a local Dutch newspaper. The CW:FW ratio is approximately 1:1 (46% and 54%, respectively). Although the analysis rules (described below in more detail) were primarily based on observations regarding this corpus, they also apply to other types of input text. If an input text would contain predominantly FWs, then the rules based on syntactic label sequences ensure correct demarcation of prosodic domains (5a); if no syntactic labels could be determined, then prosodic domains are derived from the CW/FW distinction (5b). Only in this latter case are problems to be expected, if subsequent unlabelled words are all classified as CW (5c).

van al wie er toen was ## blijkt hij degene te (5a)
V\FW ## V\FW
zijn ## die toch nog iets eraan heeft gedaan
V\FW ## Comp\FW
"of all who there then was ## appears he the-one to
be ## who (yet) yet something about-it has done"

de aan mijn broer verkochte ## zeer gevaarlijke auto (5b)
?\CW ## ?\FW
"the to my brother sold ## very dangerous car"

de onlangs dubbel verkochte (##) uiterst gevaarlijke auto (5c)
?\CW ?\CW ?\CW ?\CW ?\CW
"the recently doubly sold (##) extremely dangerous car"

3.2.2. The FW dictionary

The dictionary used by our algorithm was originally derived from the Dutch ULEX lexicon (Van den Broecke *et al.*, 1987). As a first approximation, words which qualified as FW (considering their syntactic label in ULEX) were selected. Thus, the dictionary contained all Dutch articles ($n=5$), prepositions ($n=60$), pronouns ($n=80$), conjunctives and complementizers (pooled $n=48$). For these 193 word types, Van den Broecke *et al.* (1987) report a pooled token frequency of 36% in newspaper texts. This first dictionary was modified in several ways:

- Adverbial FWs were added (e.g. **daar, er** "there").
- In Dutch, new complementizers (**waar** + Prep) and adverbs (**daar** + Prep, **er** + Prep, **hier** + Prep) can be constructed. Not all combinations of the "root" with a preposition were included in the ULEX lexicon. If such a word occurred regularly in our newspaper text corpus, then it was included in the dictionary.
- For verbal FWs, the infinitive, participle, and all inflected (past and present) forms were added. A total of 31 verbs were considered as semantically empty, and hence, as FWs.
- CWs were included in the dictionary (with their appropriate syntactic label), if they behave anomalously with regard to phrasing and/or accentuation (e.g. because they are part of an idiomatic expression in which no prosodic boundary may be interposed, or because the CW may not be accented).
- Particles of separable verbs (e.g. **toe** as in **toe + zenden**) were stored with a new syntactic label (*satellite*), unless the particle was already stored as a preposition (e.g. **op** as in **op + bellen**).
- For some words, the syntactic label was adjusted, in order to arrive at a labelling which was more relevant for our purposes. For example, the label of **beide** "both" was changed from pronoun to numeral, since this word usually functions as a numeric quantifier in NPs.

In addition to the syntactic and prosodic labels, additional features may be specified. These features trigger-specific analysis rules. Thus, the rules themselves do not refer to word strings, but only to features attributed to these word strings.

3.2.3. Word labelling

If an input word is found in the dictionary, then its syntactic label(s) is (are) copied from the dictionary. Words which are not found in the dictionary are given the prosodic label CW. Subsequent syntactic labelling concentrates on these CWs. As reported in Section 3.2.1 above, approximately 54% of the word tokens in our newspaper text corpus were found as FW in the dictionary.

First, syntactic labels are generated on the basis of formal properties of the CW string. In the future, the syntactic label(s) for each word will be provided by a separate morphological parser (Baart & Heemskerk, 1988; Heemskerk, 1989), rather than by the following rules of thumb. At this moment, the generation of syntactic labels is triggered by e.g. affixes and orthographic conventions. For example, words ending in **-dt** must be inflected verbs (**het brandt** "it burns"); words ending in **-en** are either verbs (plural inflected or infinitive) or plural nouns (**zij bakken** "they fry", **de bakken** "the trays"), and words ending in e.g. **-zaam**, **-lijk**, **-ig** must be adjectives or adverbs.

Second, words with multiple syntactic labels (either words with labels from the dictionary, or CWs with multiple labels generated by rule) must be disambiguated. Once again, this is done on the basis of restrictions on label sequences: linguistically motivated disambiguation can only be achieved when taking account of the context.

This selection of the syntactic labels generated is often based on linguistic constraints on word sequences: for example, an inflected verb may not be preceded by an article (since the article label was found in the dictionary, the non-verb label should be selected for the second word). In addition, several rules employ statistical constraints: for example, prepositions are usually followed by nouns, rather than by verbs. Such probabilistic observations are based on our analyses of the newspaper text corpus. Although these rules filter many incorrect labels [as in (6a)], they may occasionally introduce errors, as in (6b):

vogels broeden **in nesten** "birds breed in nests" (6a)
hij wil zich daar **in mengen** "he wants (to) himself there in mingle" (6b)

In addition, orthographic conventions guide the selection of syntactic labels: a word containing a hyphen is a compound, hence labeled as noun; strings containing digits are labeled as numeral; words starting with a capital (not sentence-initial) are proper names rather than verbs; words longer than 13 characters are probably nouns (e.g. **wapenhandelaren** "weapon-traders") rather than long verb strings (a verb like **herprogrammeren** "reprogram" is relatively rare), etc.

In the future, the disambiguation rules are to be refined and extended, since the morphological parser will generate a greater number of multiple syntactic labels, for a greater number of words, as compared to these orthographic rules of thumb described above.

3.2.4. Prosodic domains

In the PSS, lower prosodic domains are strictly enclosed within higher domains: they cannot straddle the boundaries of higher domains ("strict layer hypothesis"; Nespor & Vogel, 1986). In order to achieve this hierarchy, our algorithm starts by demarcating the higher Int domains. Subsequently, Phi domains are demarcated within these Int domains. Note that this procedure deviates from the theoretical construction of prosodic domains (e.g. Nespor & Vogel, 1986).

Int domains. Several rules demarcate Int domains by identifying subordinate clauses. As explained before, these rules insert an Int boundary between two adjacent words which cannot both belong to the same clause (as may be deduced from their syntactic and prosodic labels). At this point, it must be noted that our object language, Dutch, is an SOV language, where the inflected verb takes the final position in subordinate clauses, and the second position in main clauses. English, by contrast, is an SVO language. Consequently, some generalizations in our rules are allowed in Dutch, while they do not apply to English.

Below follow some examples of rules inserting Int boundaries, on the basis of syntactic word labels. Rule (7) demarcates the end of a subordinate clause (with final verb). Rule (8) is an example of a rule demarcating the beginning of a subordinate clause, while (9) demarcates two juxtaposed sentences; it is determined from the right-hand context of the conjunctive word whether this word links two subclauses (and not two other constituents). The colon indicates a subclassification; the English glosses illustrate that some of these rules can also be applied in English.

0→IntBound/ <VERB:INFL> _ <VERB:INFL> (7)

- omdat het geen haast **had ## deed** ik het later
  "because it no hurry had ## did I it later"
- de fresco's die hij **schilderde ## zien** er nog fris en kleurig uit
  "the frescoes which he painted ## look still fresh and colourful"

0→IntBound/ <VERB:PART> _ <NOT<VERB> > (8)

- hij is erin **geslaagd ## om** dit goed weer te geven
  "he has (in-it) succeeded ## this correctly to represent"
- hij heeft **beloofd ## morgen** te komen
  "he has promised ## tomorrow to come"

0→IntBound/ _ <CONJ> {<VERB:INFL>} (9)
{<FW>}

- hij raapte een steen op **## maar aarzelde** om die weg te gooien
  "he picked up a stone ## but hesitated it away to throw"
- Thero excuseerde zichzelf **## en hij** verliet de kamer
  "Thero excused himself ## and he left the room"

In addition, there are also rules which demarcate complex sentence-initial constituents (mostly subject NPs) as separate Int domains. Other rules delimit Int domains on the basis of orthographic punctuation (comma's, quotes and parentheses). Some caution is required, however, because not all instances of these puntuation signs correspond to Int boundaries:

is het geen kwestie van (taal)tolerantie om te kiezen voor 'jij'? (10)
"is it not a matter of (language) tolerance to choose 'you'?"

In this case, the (string between) parentheses should be ignored, while the quotes may trigger the labelling of the pronoun **jij** as a noun, as well as accentuation of this word.

Phi domains. For the demarcation of Phi domains, the same principle is applied as in the case of Int domains: insert a Phi boundary between two adjacent words which cannot both belong to the same Phi domain. The most important of these rules uses the fact that the prosodic head of a Phi is always its rightmost lexical item. Usually, this head is a CW. FWs always belong to the prosodic head on their right-hand side, i.e. FWs belong to the right-hand Phi domain. Hence, a Phi boundary may be assumed between a CW (which is the prosodic head of the left-hand Phi domain) and a following FW (which is a specifier of the right-hand prosodic head):

0→PhiBound/ <CW> _ <FW> (11)
• she **prunes** # **the** red **roses** # **in** her garden

Of course, this requires additional rules for Phi domains whose prosodic head is not a CW, but rule (11) correctly identifies the majority of Phi domains.

Another rule aims at demarcating verbal clusters (i.e. predicates) as separate Phi domains. This is motivated by: (1) the fact that the verbal cluster may separate subject and object arguments, corresponding to separate prosodic domains; and (2) the special accentuation behaviour of verbal clusters (see Section 3.3.2 below).

The above rules for Phi boundary insertion leave long strings of non-verbal CWs intact, i.e. such strings are not divided into multiple Phi domains. Of course, this may be incorrect in cases where the CWs belong to different thematic constituents, such as (12). Therefore, another rule splits these CW strings, by inserting a prosodic boundary before an adverb, which is usually *not* a specifier to its left-hand neighbour.

ik drink # over het algemeen # **liever** sterke koffie (12)
FW FW N\CW Adv\CW Adj\CW N\CW
"I drink # in general # preferably strong coffee"

Finally, some of the resulting Phi domains (e.g. those containing only an FW prosodic head) are merged with their left-hand neighbour.

3.3. From PSS to prosody

In the second part of our algorithm, abstract prosodic phenomena are derived from the PSS constructed by the first part. At this moment, the PSS is used only to determine phrase boundaries and accentuation, as mentioned in Section 1. In theory, however, other aspects of the output sentence prosody (e.g. phonological sandhi phenomena) can also be derived from the PSS.

3.3.1. Phrasing

At this moment, the PSS is used directly to split the input sentence into separate phrases. Each resulting Int boundary becomes manifest as a prosodic break, which is realized in the output synthetic speech as a pause (250 ms) accompanied by an apropriate F_0 movement ('t Hart & Collier, 1975; Terken & Collier, 1989). This procedure, however, results in disfluency in the synthetic speech output, thus inhibiting (rather than facilitating) its correct perception. Apparently, too many breaks are present in the speech stream, and the phonetic means by which they are realized may be too strong. Currently, two solutions are investigated to improve this situation.

First, the *number* of Int boundaries can be reduced, by means of the restructuring of Int domains (as mentioned in Section 2). In other words, two adjacent Int domains are collapsed into a single one, by rules deleting the intermediate Int boundary. Again, this restructuring should be guided by theoretical considerations. Shorter Int domains (in terms of number of words and syllables) are more prone to collapsing; the resulting Int domains should be of approximately equal length; restructuring depends (to a certain extent) on the syntactic weight of the Int domains involved.

Second, the *phonetic realization* of Int boundaries as prosodic breaks in the output speech could be differentiated, depending on the weight of the Int boundary, as well as on the length and function of the corresponding Int domains. Relatively weak Int boundaries may be realized by phonetic means which are perceptually less salient, e.g. prepausal lengthening rather than F_0 movements.

3.3.2. Accentuation

A fitting contour of a sentence can be derived automatically (at least, in the case of Dutch), if phrase boundaries and accent locations are known ('t Hart & Collier, 1975; Van Wijk & Kempen, 1985; Terken & Collier, 1989). In this section, we will describe how accents can be derived from the PSS. The result is an abstract word property: either plus or minus *accent*. This property becomes manifest primarily by intonational means, but also in segmental durations, intensity, vowel reduction, etc. (Nooteboom & Kruyt, 1987; see Section 1).

The accentuation component of our algorithm attempts to imitate several theoretical aspects which are known to affect accentuation. After reviewing these factors, our approach in accentuation is discussed.

Theory. Evidently, FWs seldom receive accent. It is a consequence of their FW status that they have a reduced prominence, and are usually left unaccented (Kruyt, 1985).

Second, in Dutch, the prosodic head (i.e. last CW in Phi domain) is the word which receives *integrative accent*. An accent on this prosodic head lends prominence (or "focus") to the whole domain, and not only to the carrier word (Baart, 1987).

Third, discourse context, specifically the distinction between "given" and "new" information, strongly influences accentuation (Fuchs, 1984). A domain must be unaccented if it refers to information which is (supposed to be) already given to the listener ("known", "old"). Usually, only domains introducing new information are accented (by means of an integrative accent), as demonstrated below (where accent is indicated by capitals):

he came # by **CAR** (13a)
his **car** # was BLUE (13b)

Kruyt (1985) and Terken (1985) have confirmed the role of this distinction between "given" and "new" in accentuation. Apparently, listeners rely on this distinction for their understanding of the spoken utterance (Terken & Nooteboom, 1987).

Fourth, the thematic relations between prosodic domains play an important role (Gussenhoven, 1984; Baart, 1987). As an example, consider the accentuation of the predicate (i.e. prosodic domain corresponding to the main verb or verbal group). Even if it refers to new information, this word (group) is usually not accented (14). Under two conditions, however, the predicate must be accented.

- If the predicate is not adjacent to (any of) its argument(s); this may happen if a non-argument constituent (sentence modifier, adverbial phrase) intervenes between the predicate and its arguments, as in (15).
  The relevance of the modifier status of the intervenient can be inferred from comparing **met de bloemen** (14b) and **met de machine** (15). Neither are arguments to the predicate, but only the latter one is a sentence modifier, while the former one depends syntactically on **het gazon**, yielding a complex argument.

- If all of the arguments of the predicate are unaccented; this may happen if the arguments convey given information (16a), or if they contain only FWs (16b).

ik heb # het GAZON # gemaaid (14a)
"I've # the lawn # mown"
ik heb # het gazon # met de BLOEMEN # gemaaid (14b)
"I've # the lawn # with the flowers # mown"

ik heb # het GAZON # met de MACHINE # GEMAAID (15)
"I've # the lawn # with the machine # mown"

(hoe is het met het gazon?) ("how is the lawn?")
ik heb # het gazon # GEMAAID "I've # the lawn # mown" (16a)
ik heb # het # GEMAAID "I've # it # mown" (16b)

Finally, rhythmic factors influence accentuation. The occurrence of multiple adjacent accents is avoided. In such cases, one of the accents is removed or shifted to a different word (see Kager & Visch, 1988; Visch, 1989 for a lengthier discussion on this issue).

Algorithm. The accentuation rules in the algorithm attempt to imitate the theoretical account of accentuation discussed above. In addition, the rules refer to the PSS, as well as to the syntactic and prosodic word labelling.

First, CWs are accented. Two types of words are excluded from this accentuation, viz. *verbs* and semantically empty words (e.g. **gulden** "guilder"). In addition, some idiosyncratic words are also accented (e.g. **nooit** "never"). Both types of anomalous words are listed in the dictionary. This procedure generates too many accents. Therefore, subsequent rules strip CWs of their accent, under conditions where an accent is known to be wrong: (1) rhythmic deaccentuation, (2) deaccentuation of words conveying "given" information. In addition, (3) verbs (or verb groups) are accented in some specific contexts.

Rhythmic de-accentuation. In order to obtain rhythmic accentuation patterns, the middle one of three adjacent accented CWs is de-accented. Both the PSS and the syntactic word labelling are taken into account: all three words must belong to a single Phi domain, and they must belong to certain sequences of syntactic labels:

QuantNum	*X*	*Noun*		(17)
drie	duizend	boeken	"three thousand books"	
zeer	lage	temperatuur	"very low temperature"	

Adverb	*Adj*	*Noun*		(18)
evenwijdig	lopende	spoorlijnen	"parallel running tracks"	
zeer	verbaasde	toeschouwers	"very amazed spectators"	

These conditions are stricter than one would expect from a theoretical viewpoint, in order to avoid incorrect over-applications. In the future, more contexts could be added to trigger rhythmic de-accentuation.

De-accentuation of "given" information. Words which can safely be assumed to convey given information, are also de-accented. Although the scope of our algorithm is limited to a single sentence of the input text, the rules can nevertheless infer whether words in the sentence under analysis have occurred before. Deictic qualifiers (such as **dit, deze** "this", **dergelijke, zulke** "such") imply that the following term refers to information which has already been introduced ("given"). Any words following this cue word in the same Phi domain are de-accented, as indicated by (19):

he came # by CAR (19a)
he has BORROWED # **this vehicle** # from a FRIEND (19b)

Two groups of de-accenting ("given") qualifiers are distinguished, for reasons related to verb accentuation. Both groups trigger de-accentuation of subsequent words within the Phi domain in which they occur. The second group includes all inflected comparative forms of adjectives (i.e., all words ending in **-ere**). Obviously, the basis of such rules (as well as of the "given" qualifier sets) is probabilistic. Errors occur in a minority of cases, but examples like (20) certainly represent the majority:

het GEVOLG # van **zulke** temperaturen # is VRESELIJK (20a)
"the consequence # of such temperatures # is awful"
ONDERZOEK # heeft # de **hogere** sterfkans # BEWEZEN (20b)
"research # has # the higher death risk # proven"

In addition, de-accentuation of given information is applied in a few specific contexts: epitheta before a proper name (as in **koningin Beatrix** "queen Beatrix"), sentences following (in)direct speech quotation, and clause-medial unit measures following a numeral.

Verb accentuation. Under certain conditions, the predicate (verb group) must be accented, as discussed in the Theory subsection above. This is required if a sentence modifier intervenes between the verb (group) and its arguments, or if all of the arguments are unaccented (not in focus)—or simply absent.

The rules which accentuate a verb work on single CW verbs, as well as on longer sequences of verbs, of which the first CW is accented. In addition, the rules work on complex verbs, of which the stem and particle can occur separately. Stem and particle occur together only in particles (21a; where Dutch orthography requires that the two parts are written as a single word). But in infinitive constructions, the infinitive marker **te** can be interposed (21b), whereas in inflected forms in Dutch, the inflected stem (in second position in main clause) may be far away from the particle (in the "original" clause-final position), as in (21c). An interesting property of these separable verbs is that the *particle* is accented, even if the stem is far away (21b,21c):

ik **heb** # mijn MOEDER # VANDAAG # **OPGEBELD** (21a)
"I've # my mother # today # phoned"
ik heb GEPROBEERD # mijn MOEDER # VANDAAG # **OP te bellen** (21b)
"I've tried # my mother # today # to phone"
ik **belde** # haar # VANDAAG # **OP**
"I phoned # her # today # -" (21c)

The accentuation of these separable verbs is identical to other verbs, and incorrect non-accentuation of the *particle* part is accordingly a serious error. However, particles are often identical to prepositions [e.g. the ambiguous **op** in (21)]. This presents difficulties in detecting verbal particles. Some can be identified because they occur immediately before the infinitive marker **te** (21b); others may be identified because they occur in clause-final position (21c), hence cannot introduce a PP, and hence cannot be prepositions. The rules to be discussed below include these stranded clause-final particles in their focus, so that the particles can carry the accent of separable verbs.

The verb (group) itself must be followed by an Int (or sentence) boundary. This guarantees that the verb is syntactically clause-final and not clause-medial, where it would typically be followed by (and hence, be adjacent to) an argument. Adverbial phrases consisting of PPs are detected through their initial preposition. That is, PPs starting with e.g. **ondanks** "notwithstanding", **in** "in" (followed by proper name), **sinds** "since" are used as sentence modifiers (22b), rather than as constituents within a complex NP.

hij heeft # het GAZON # **VANDAAG**$_{\mathbf{Adv}}$ # **GEMAAID** (22a)
"he has # the lawn # today # mown"
hij heeft # het GAZON # **ondanks** de REGEN # **GEMAAID** (22b)
"he has # the lawn # in-spite-of the rain # mown"

If the predicate is accompanied by an unaccented argument, then the predicate (verb group) must be accented. Such arguments are indicated by their deictic "given" qualifiers (See De-accentuation of "Given" Information subsection above). However, only the first group of these qualifiers triggers verb accentuation (23). The (second) group including inflected adjectives frequently takes the integrative accent in an argument + predicate construction, as in (24).

de ZEEHONDEN # hebben # **deze** lage temperature # **OVERLEEFD** (23)
"the seals # have # these low temperatures # survived"

ze heeft # de **HOGERE** weg # **genomen** (24)
"she has # the higher road # taken"

Likewise, the verb (group) is accented if its (object) argument is a pronoun (unaccented FW), or if there is no object argument at all. The latter condition can be assessed from the fact that the verb is clause-final, and preceded by a single Phi domain (25a) which does not contain any pronouns [cf. (25b)].

de zeehonden # zijn **GESTORVEN** (25a)
"the seals # have died"
dat **ik mijn** ZUSTER # heb **ontmoet** (25b)
"that I my sister # have met"

4. Evaluation

As explained in Section 1, the algorithm described in the previous section aims at establishing the correct abstract sentence prosody (accentuation and phrasing). It is assumed that these aspects (or rather, the appropriate phonetic correlates of these aspects) make synthetic speech more natural and intelligible. In order to evaluate whether our algorithm achieves this aim, two methods are possible. First, we can compare its output with the accentuation and pausing as produced by a human speaker. Ideally, there should be no difference between the two; any differences found should not disturb the semantic and pragmatic equivalence between the natural and synthetic versions. (One could discriminate between obligatory and optional accents and pauses; differences should be limited to the optional accents and pauses.) Second, the output can be evaluated from a perceptual viewpoint, viz. by investigating whether synthetic speech is sufficiently intelligible and adequate, if accents and pauses are derived automatically.

4.1. Comparison with natural prosody

In order to evaluate the algorithm, a comparison was made between its abstract prosody and natural speech. The latter was produced by a professional speaker, who read aloud several texts, the majority of which were originally written as radio news bulletins (and hence, meant to be read aloud). In total, the material consisted of 10 766 words [5273 CWs, 5493 FWs] in 600 sentences, grouped into 43 texts. Recordings were made at the Institute for Perception Research (IPO) in Eindhoven, The Netherlands. The recordings were transcribed with respect to accents and prosodic boundaries by the two authors (and occasionally, by a third transcriber). If all transcribers agreed, then a word was considered to be accented, otherwise not. The results in Table I and II below show the degree of convergence (agreement) between the human speaker and the algorithm PROS, with respect to phrasing and accentuation. These results are only of limited value, however, since a considerable amount of variation is allowed with respect to abstract prosody. A difference in phrasing and/or accentuation between man and algorithm does not necessarily imply that the latter has been wrong. The sentence pair (26) below illustrates this variation: both versions are equally acceptable, and roughly equivalent. All cases of discrepancies (between the output and naturally produced prosody) where the two versions are different and yet equivalent, were marked. The tables below also give the remaining number of discrepancies (in parentheses), if these cases are discarded.

een aantal ONDERZOEKERS meent overigens ## dat de VRAAG ## (26a)
of passief meeroken SCHADELIJK is ##
al LANG positief kan worden BEANTWOORD
een AANTAL onderzoekers MEENT overigens ## dat de vraag (26b)
of PASSIEF MEEROKEN SCHADELIJK is ##
al lang POSITIEF kan worden beantwoord
"a number (of) researchers thinks indeed ## that the question (##)
whether passive smoking injurious is ##
already positively can be answered"

The data in Table I show that the algorithm predicts 65% of the "human" prosodic boundaries correctly, and that the same decision is taken in 90% of all relevant cases (error rate 10%). This result remains about the same if we ignore those prosodic boundaries which are taken from orthographic punctuation in the input text (e.g. commas and parentheses; $n = 550$; we assume that these were also realized by the human speaker). If different but equivalent phrasings are discarded, however, then performance

TABLE I. Comparison between the numbers of (possible) prosodic boundaries, as realised by a human speaker and by the PROS algorithm. Differences yielding equivalent phrasings are discarded in the adjusted numbers (between parentheses; see text). In total, 10 766 (word boundaries) minus 599 (sentence boundaries) = 10 167 (intra-sentence word boundaries) were compared

		Human speaker		
		Realized	Not realized	Total
PROS	Realized	855 (1138)	503 (282)	1358
	Not realized	464 (181)	8345 (8566)	8809
	Total	1319	8848	10 167

of our algorithm increases considerably: 86% of the naturally produced boundaries are predicted, and the error rate drops to 5%.

The data with regard to accentuation (Table II) show an agreement of 85% if all words are pooled. In this connection, it must be noted that the two error types are not independent. Accents usually occur in a rhythmic pattern. Therefore, incorrect accentuation of one word corresponds to an incorrect non-accent on a neighbouring word, as in the fragments **een aantal onderzockers meent** and **al lang positief** in (26) above. The different accentuations of such fragments are not equivalent, since semantic prominence is guided towards different words. Hence, only few discrepancies yield equivalent accentuations, and may, therefore, be discarded as errors (yielding 87% agreement).

TABLE II. Comparison between the numbers of accented words as produced by a human speaker and by the PROS algorithm. Differences yielding equivalent accentuations are discarded in the adjusted numbers (between parentheses; see text)

		Human speaker		
		+Acc	−Acc	Total
PROS	+Acc	3484 (3533)	883 (810)	4367
	−Acc	689 (640)	5710 (5783)	6399
	Total	4173	6593	10 766

Since FWs are seldom accented, the agreement in (non-)accentuation is considerably higher for FWs (viz. 94%; this class includes all particles) as compared to CWs (77%).

4.2. Perceptual evaluation of accentuation

Van Bezooijen (1989) has evaluated the output of the algorithm from a perceptual perspective. Eight texts (total 24 sentences) and eight isolated sentences were fed into the algorithm. The output abstract prosodic markers were converted into phonetic prosody (most notably, pitch accents) in diphone speech. In addition, three control conditions were created:

1. *random*: the same number of accents as produced by the algorithm ($n = 274$) were distributed at random over the CWs in the text (low control).
2. *subjects*: beforehand, subjects were asked to indicate the accentuation which they considered to be optimal; a word in the stimulus material was accented if 7 out of 12 subjects agreed (high control 1).
3. *natural*: a word was accented if the professional human speaker (viz. the same as mentioned in Section 4.1) had accented that word, as agreed by three out of four transcribers (high control 2).

These four accent versions of each sentence were then presented to 12 listeners. Their task was to rate the accentuation of each sentence on a 10-point adequacy scale. Results are summarized in Table III below. From these results, Van Bezooijen (1989) concludes that the algorithm produces sufficiently adequate accentuation, although its output is still defective in several respects.

TABLE III. Mean scores on a 10-point adequacy scale, averaged over 32 sentences and 12 listeners, for four accentuation conditions

Random	PROS	Subjects	Natural
4·6	6·0	7·7	7·4

4.3. *Error analysis*

Both in phrasing and accentuation, two types of errors were discriminated above: (1) "misses"; where the algorithm failed to predict a phrase boundary or accent observed in natural speech; (2) "false alarms"; where the algorithm produced a phrase boundary or accent which is absent in natural speech. Errors were classified by the first author, according to the responsible factor. Results are summarized in Tables IV and V below.

TABLE IV. Number of unpredicted ("miss"), uncorrectly predicted ("false alarm") and total number of differing phrase boundaries, in five error categories, as observed in comparison between algorithm output and corresponding natural speech (see text). Discrepancies yielding equivalent phrasings are excluded from marginal totals

Category	"Miss"	"False alarm"	Total
[Equivalent]	283	221	
Label error	3	28	32
Syntax error	151	187	338
Phrase length error	8	32	40
Punctuation/textual division error	4	14	18
Other	15	21	46
Total	181	282	463

These results show that most *phrasing* errors (73%) are caused by incorrect mapping from syntactic constituents to prosodic domains. Most often, subordinate clauses are incorrectly demarcated, as in (27) (where "/" denotes a missing Int boundary):

de strepen worden doorgesneden ## langs de deuren #/ (27)
om het instappen # te vergemakkelijken
"the strips are cut ## along the doors #/
in-order-to the getting-in # to facilitate"

Such errors are obviously serious, since they result in "garden path" sentences: the prosody signals an incorrect syntactic structure. Some of these errors in syntactic demarcation and phrasing may also propagate into subsequent accentuation errors ($n = 69$), see Table V).

TABLE V. Number of unpredicted ("miss"), uncorrectly predicted ("false alarm") and total number of differing accents (accented words), in 10 error categories, as observed in comparison between algorithm output and corresponding natural speech (see text). Discrepancies yielding equivalent accentuations are excluded from marginal totals

Category	"Miss"	"False alarm"	Total
[Equivalent]	49	76	
Dictionary error	27	73	100
Label error	38	57	95
Phrase error	2	11	13
Label + phrase error	5	11	16
Syntax error	7	33	40
Idiomatic expression	12	62	74
Semantic/pragmatic/focus	507	492	999
Rhythmic error	1	12	13
Accent error	27	52	79
Other	14	4	18
Total	640	810	1447

Most errors in *accentuation* (69%) are caused by semantic, pragmatic and contextual factors. Any accentuation algorithm should decide which words convey "given" information, and adjust the accentuation accordingly. Moreover, it should also decide which words are important, and then assign a contrastive accent on these words (or an integrative accent on the nucleus word). Our algorithm performs poorly on both of these tasks. Its poor performance on the first task is primarily due to the fact that its scope is limited to a single sentence, yielding errors such as (28):

de OLIEVLEK # zal aan DUIZENDEN VOGELS # het LEVEN kosten (28a)
er zijn al # TIENTALLEN **VOGELS** # DOOD AANGETROFFEN (28b)
"the oil-slick # will of thousands [of] birds # the life cost"
"there are already # tens [dozens of] birds # dead found"

The obvious remedy might be to widen this scope, e.g. by maintaining a buffer of content words (Silverman, 1987) or root morphemes (Hirschberg, 1990) across sentences, and resetting this buffer at paragraph boundaries. This strategy would avoid some errors, but co-references involving synonyms and paraphrases are likely to go undetected—although these are more frequent than the former type.

Moreover, there is no perfect correspondence between the distinctions "given"/"new" and ± accent. Words conveying "new" information can be unaccented (29a), and (implicitly) "given" words can be accented (29b):

de HERENIGING # **van de twee Duitslanden** (29a)
"the reunion # of the two Germanies"
ZIJN salaris # is VEEL hoger # dan het UWE (29b)
"his salary # is much higher # than yours"

In the latter case, contrastive accents are selected solely on pragmatic and contextual grounds. Such accents can be derived partly from statistical regularities and other heuristics (Monaghan, 1990*c*; note that such heuristics may sometimes yield inappropriate results, as in our "dictionary errors"). Yet, the majority of such errors cannot be solved without a better understanding of the complex relation between accentuation and our knowledge of the real world.

The remaining discrepancies result from errors during the derivation process (labeling, syntactic constituency, phrasing, phrase length adjustment, accentuation). Some of these were solved after minor improvements in our rule set. Finally the relatively low number of discrepancies due to labeling errors is worth noting ($n=32$ in phrasing, $n=111$ in accentuation). This suggests that our primitive means to determine syntactic word class may be sufficient for the purpose of sentence prosody.

5. Conclusion

In the present paper, it was argued that suprasegmental phenomena in synthetic speech should be derived, via abstract accentuation and phrasing, from an underlying linguistic sentence representation. We, therefore, attempted to establish the prosodic sentence structure ("PSS"; Nespor & Vogel, 1982, 1986). This structure is based on a well-established phonological theory (supported by independent evidence from various languages), and predicts many prosodic phenomena. Yet, it does not require exhaustive syntactic analysis of the input sentence, which (even if required at all) is currently unavailable.

The PSS is approximated on the basis of syntactic word labelling. Subsequently, phrasing and accentuation are derived from the PSS, while for accentuation some semantic and contextual information is also taken into account.

A large proportion of the errors in the resulting phrase boundaries will be solved with more advanced syntactic parsing. Accentuation errors are mainly caused by contextual and pragmatic factors. It is the nature of things, however, that such errors cannot be avoided in a principled way. Fortunately, human speakers are free to produce utterances (and prosodic patterns) which do not adhere to linguistic rules.

This research was supported by the Dutch research programme for "Analyse en Synthese van Spraak" (SPIN/ASSP). We thank Ad Aerts (MediaSystemen, Haarlem, The Netherlands) for supplying newspaper text corpora; Sidonne Bos, Eldrid Bringmann, Marieke van Capellen, Yvonne van Holsteijn, Mariken ter Keurs, Jeroen Reizevoort and Tom Veenhof for accurately performing a great variety of tedious tasks; and Vincent van Heuven, Bert Schouten, and an anonymous reviewer for their valuable comments on the manuscript.

References

Allen, J., Hunnicutt, M. S. & Klatt, D. (1987). *From Text to Speech: The MITalk System*. Cambridge University Press, Cambridge.

Baart, J. L. G. (1987). Focus, syntax, and accent placement: towards a rule system for the derivation of pitch accent patterns in Dutch as spoken by humans and machines. Dissertation, Leiden.

Baart, J. L. G. & Heemskerk, J. S. (1988). The problem of ambiguity in morphological analysis for a Dutch text-to-speech system. In *Proceedings 7th FASE Symposium (SPEECH '88)*. Edinburgh. Volume 3, pp. 959–965.

Bachenko, J. & Fitzpatrick, E. (1990). A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, **16**, 155–170.

Bailly, G. (1989). Integration of rhythmic and syntactic constraints in a model of generation of French prosody. *Speech Communication*, **8**, 137 46.

Carlson, R., Granström, B. & Hunnicutt, S. (1989). Multilingual Text-to-speech Development and Applications. Internal report, Department of Speech Communication and Music Acoustics, Royal Institute of Technology (KTH), Stockholm.

Collier, R. & 't Hart, H. (1975). The role of intonation in speech perception. In *Structure and Process in Speech Perception* (Cohen, A. and Nooteboom, S. G., eds), pp. 107–21. Springer, Berlin.

Cooper, W. E. & Paccia-Cooper, J. (1980). *Syntax and Speech*. Harvard University Press, Cambridge, Massachussetts.

Cooper, W. E. & Sorensen, J. M. (1977). Fundamental frequency contours at syntactic boundaries. *Journal of the Acoustical Society of America*, **62**, 683–692.

Cutler, A. (1982). Prosody and sentence perception in English. In *Perspectives on Mental Representation: Experimental and Theoretical Studies of Cognitive Processes and Capacities* (Mehler, J., Walker, E. C. T. and Garrett, M., eds), pp. 201–216. Lawrence Erlbaum, Hillsdale, New Jersey.

Cutler, A. & Clifton, C. (1984). The use of prosodic information in word recognition. In *Attention and Performance. Volume X: Control of Language Processes* (Bouma, H. and Bouwhuis, D. G., eds), pp. 183–196. Lawrence Erlbaum, London.

Fuchs, A. (1984). 'Deaccenting' and 'default accent'. In *Intonation, Accent and Rhythm* (Gibbon, D. and Richter, H., eds). De Gruyter, Berlin.

Gee, J. P. & Grosjean, F. (1983). Performance structures: a psycholinguistic and linguistic appraisal. *Cognitive Psychology*, **15**, 411–458.

Goldman-Eisler, F. (1972). Pauses, clauses, sentences. *Language and Speech*, **15**, 114–21.

Gussenhoven, C. (1984). *On the Grammar and Semantics of Sentence Accents*. Foris, Dordrecht.

Heemskerk, J. (1989). Morphological parsing and lexical morphology. In *Linguistics in the Netherlands 1989* (Bennis H. and Van Kemenade, A., eds), pp. 61–70. Foris, Dordrecht.

Hirschberg, J. (1990). Using discourse context to guide pitch accents decisions in synthetic speech. In *Proceedings of the ESCA Workshop on Speech Synthesis*. Autrans, pp. 181–184.

Kager, R. & Visch, E. (1988). Metrical constituency and rhythmic adjustment. *Phonology*, **5**, 21–71.

Kerkhoff, J., Wester, J. & Boves, L. (1984). A compiler for implementing the linguistic phase of a text-to-speech conversion system. In *Linguistics in the Netherlands 1984* (Bennis, H. and Van Lessen Kloeke, W. U. S., eds), pp. 111–117. Foris, Dordrecht.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, **3**, 129–140.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, **59**, 1208–1221.

Klatt, D. H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, **82**, 737 793.

Koopmans-Van Beinum, F. J. (1980). Vowel contrast reduction: an acoustic and perceptual study of Dutch vowels in various speech conditions. Dissertation, Amsterdam.

Kruyt, J. G. (1985). Accents from speakers to listeners: an experimental study of the production and perception of accent patterns in Dutch. Dissertation, Leiden.

Kulas, W. & Rühl, H.-W. (1985). Syntex—unrestricted conversion of text-to-speech for German. In *New Systems and Architectures for Automatic Speech Recognition and Synthesis* (De Mori, R. and Suen, C. Y. eds), pp. 517–335. Springer, Berlin.

Ladd, D. R. (1987). A model of intonational phonology for use in speech synthesis by rule. In *Proceedings European Conference of Speech Technology*. Edinburgh, Volume 1, pp. 21 24.

Lehiste, I. (1970). *Suprasegmentals.* The MIT Press, Cambridge, Massachussetts.

Monaghan, A. I. C. (1990*a*). Rhythm and stress-shift in speech synthesis. *Computer Speech and Language*, **4**, 71–78.

Monaghan, A. I. C. (1990*b*). A multi-phase parsing strategy for unrestricted text. In *Proceedings of the ESCA Workshop on Speech Synthesis.* Autrans, pp. 109–112.

Monaghan, A. I. C. (1990*c*). Treating anaphora in the CSTR text-to-speech system. In *Proceedings of the ESCA Workshop on Speech Synthesis.* Autrans, pp. 113–116.

Nespor, M. & Vogel, I. (1982). Prosodic domains of external sandhi rules. In *The Structure of Phonological Representations* (Van der Hulst, H. and Smith, N., eds), Volume 1, pp. 225–255. Foris, Dordrecht.

Nespor, M. & Vogel, I. (1986). *Prosodic Phonology.* Foris, Dordrecht.

Nooteboom, S. G. (1985). A functional view of prosodic timing in speech. In *Time, Mind, and Behavior* (Michon, J. A., ed), pp. 242–252. Springer, Berlin.

Nooteboom, S. G., Brokx, J. P. L. & De Rooij, J. J. (1978). Contributions of prosody to speech perception. In *Studies in the Perception of Language* (Levelt, W. J. M. and Flores d'Arcais, G. G., eds), pp. 75–107. John Wiley, Chichester.

Nooteboom, S. G. & Kruyt, J. G. (1987). Accents, focus distribution, and the perceived distribution of given and new information: an experiment. *Journal of the Acoustical Society of America*, **82**, 1512–1524.

O'Shaughnessy, D. D. (1989). Parsing with a small dictionary for applications such as text to speech. *Computational Linguistics*, **15**, 97–106.

Quazza, S., Varese, G. & Vivalda, E. (1989). Syntactic pre-processing for high quality text-to-speech. In *Proceedings European Conference on Speech Communication and Technology (EuroSpeech '89).* Paris, Volume 1, pp. 506–509.

Scharpff, P. J. & Van Heuven, V. J. (1988). Effects of pause insertion on the intelligibility of low quality speech. In *Proceedings 7th FASE Symposium (Speech '88).* Edinburgh, Volume 1, pp. 261–268.

Selkirk, E. O. (1984). *Phonology and Syntax: the Relation Between Sound and Structure.* The MIT Press, Cambridge, Massachussetts.

Selkirk, E. O. (1986). On derived domains in sentence prosody. In *Phonology Yearbook* (Ewen, C. J. and Anderson, J. M., eds), Volume 3, pp. 371–405.

Silverman, K. (1987). The structure and processing of fundamental frequency contours. PhD thesis, Cambridge University.

Terken, J. M. B. (1985). Use and function of intonation: some experiments. Dissertation, Leiden.

Terken, J. M. B. & Collier, R. (1989). Automatic synthesis of natural-sounding intonation for text-to-speech conversion in Dutch. In *Proceedings European Conference on Speech Communication and Technology (EuroSpeech '89).* Paris, Volume 1, pp. 357–359.

Terken, J. & Nooteboom, S. G. (1987). Opposite effects of accentuation and deaccentuation on verification latencies for Given and New information. *Language and Cognitive Processes*, **2**, 145–163.

't Hart, J. & Collier, R. (1975). Integrating different levels of intonation analysis. *Journal of Phonetics*, **3**, 235–255.

Te Lindert, E., Doedens, C. J. & Van Leeuwen, H. (1989). *Spraakmaker-1.* Stichting Spraaktechnologie, Utrecht.

Van Bezooijen, R. (1989). *Evaluation of an Algorithm for the Automatic Assignment of Sentence Accents in Written Text.* Stichting Spraaktechnologie, Utrecht.

Van den Broecke, M. P. R., Aerts, A., Reizevoort, J., Veenhof, T., Lammens, J. & Elstrodt, M. (1987). Type- and token-frequencies of wordclasses, phonemes and phoneme pairs in Dutch. *Progress Reports Institute of Phonetics Utrecht (PRIPU)*, **12**, 1–15.

Van Wijk, C. & Kempen, G. (1985). From sentence structure to intonation contour. In *Sprachsynthese: zur Synthese von natürlich gesprochener Sprache aus Texten und Konzepten* (Müller, B. S., ed.), pp. 157–182. Georg Olms, Hildesheim.

Visch, E. A. M. (1989). A metrical theory of rhythmic stress phenomena. Dissertation, Utrecht.

Wingfield, A. (1975). The intonation-syntax interaction: prosodic features in perceptual processing of sentences. In *Structure and Process in Speech Perception* (Cohen, A. & Nooteboom, S. G., eds), pp. 146–156. Springer, Berlin.