AUTOMATIC ACCENTUATION AND PROSODIC PHRASING FOR DUTCH TEXT-TO-SPEECH CONVERSION

Hugo Quené and René Kager

Research Institute for Language and Speech,
Rijksuniversiteit Utrecht, the Netherlands

ABSTRACT

Correct accentuation and phrasing improves the quality of synthetic speech. This paper discusses an algorithm which assigns both sentence accents and phrase boundaries on the basis of the prosodic sentence structure. Although this latter structure is theoretically derived from the syntactic structure, the present algorithm determines the prosodic structure by means of (linguistically and statistically motivated) rules of thumb, i.e., without exhaustive syntactic analysis. Subsequently, both prase boundaries and sentence accents are assigned on the basis of this prosodic sentence structure. In this component, several factors affecting accentuation are imitated by rules which do not directly refer to the syntactic or thematic structure.

1. INTRODUCTION

Natural speech contains many prosodic phenomena, which are related to the linguistic structure of a speech utterance. Among the most important are segmental durations, sentence accentuation, F_0 contours, and phrasing. In this paper, we will restrict ourselves to (1) sentence accentuation and (2) phrasing, which together affect the F_0 contour, pausing, and segmental durations.

Prosodic cues serve to indicate the intended linguistic structure, and thus help the listener in understanding the speech utterance correctly [3,15,16]. The perceptual role of prosodic markers becomes more important under adverse listening conditions, when the speech signal is less redundant at the segmental level. For example, synthetic speech (such as the output of a text-to-speech system) contains less phonetic cues as compared to natural speech (due to the necessary reduction of acoustic-phonetic redundance). Consequently, correct prosodic markers greatly increase the overall quality and intelligibility of synthetic speech.

In the production of natural speech, most prosodic phenomena are assumed to depend (to a varying extent) on the linguistic structure of an utterance. However, the primary factor here is not the *syntactic* sentence structure directly, but rather the related "performance structure" or *prosodic* sentence structure (hence PSS) [13,14,18,19,4]. Extending this line of thought, then, we assume that both accentuation and phrasing are related to an identical sentence structure, viz. the PSS. Note that it is also conceivable that each prosodic phenomenon has a separate 'controller structure'. The above assumption is corroborated, however, by the observation that discrepancy between two prosodic markers, viz. F_0 contour and pausing, has a negative effect on the intelligibility of synthetic speech [17].

In this paper, we will investigate whether such a relation between PSS and prosodic phenomena can be useful in text-to-speech conversion. The algorithm PROS, which is described here, aims to arrive at the correct accentuation and phrasing, on the basis of the PSS. Firstly, the PSS of an input text sentence is established, without exhaustive syntactic and pragmatic analysis. Subsequently, accentuation and phrasing are derived by means of this PSS, and output as abstract prosodic markers in the sentence representation. An earlier version of this algorithm has been described before [8]. The algorithm is intended as one of the many components in a Dutch text-to-speech system [12]. On the basis of its output, 'silence' segments are inserted in the phoneme string, and the F_0 contour is calculated [6,21].

2. PROSODIC SENTENCE STRUCTURE

In [13,14,18,19], proposals are made as to the mapping between syntactic constituents and *prosodic domains*. As is known from the phonologies of many languages, sandhi rules that apply between words have their own specific domains of application. These domains are not necessarily isomorphous to syntactic constituents. Two prosodic domains attested in this sense are the phonological phrase (Phi) and the intonational phrase (Int).

In languages such as English and Dutch, *Phi* domains consist of a lexical head (i.e., noun, verb, adverb or adjective), its left-hand specifiers, and all non-lexical words to the left up to the next lexical head. Non-lexical words are function words (hence FWs) such as prepositions, complementizers, conjunctions, and copula's. The final element (the lexical head) of each Phi is the *prosodic head*. An illustration is (1):

(1) (This sentence) (is divided) (into Int's)
(which are divided) (into Phi's)

This example clearly demonstrates that Phi domains do not necessarily correspond to syntactic constituents.

The prosodic domain *Int*, or intonational phrase, is motivated primarily by its function in phrasing (pausing, etc.). Secondly, it is the span over which intonation contours are realised. Phi and Int are in a hierarchical relation, since every Int boundary must correspond to a Phi boundary. Phi's are joined into Int's, preserving the integrity of higher-level syntactic constituents. Some constituents obligatorily form an Int, such as displaced syntactic constituents, parentheticals, and non-restrictive relative clauses. Complex NP's and subordinate clauses form Int's as much as possible, depending upon their length and upon the style of speech. An example is (2):

(2) (Int (This sentence) (is divided) (into Int's))
(Int (which are divided) (into Phi's))

Prosodic domains tend to be of equal length as much as possible, 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 one lexical head with the Phi to its left under some syntactic conditions. Very short Int's can be eliminated by merging them with adjacent Int's, and very long Int's are broken down into shorter ones.

3. PROSODIC ANALYSIS: ALGORITHM

According to phonological theory, the PSS is derived from the syntactic sentence structure, as described in the preceding section. However, this principle cannot be applied to automatic prosodic sentence analysis, for two reasons.

Firstly, there is no syntactic 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.

Even if such a parser would exist, it would run into great difficulties when analysing syntactically ambiguous sentences like the following:

(3a) ik heb (het gazon met de bloemen) (gemaaid)
I (have mowed) (the lawn with the flowers)
(3b) ik heb (het gazon) (met de machine) (gemaaid)
I (have mowed) (the lawn) (with the machine)

Solving this type of ambiguity requires a semantic and pragmatic analysis; the parser must 'know' that one cannot use flowers to mow a lawn. Again, no system exists for this type of sentence analysis.

In order to bypass these problems, the PROS algorithm attempts to derive the PSS directly, i.e. without exhaustive syntactic analysis. In other words, the prosodic domains (Phi and Int) are not established along the lines described above (viz. via an intermediate syntactic analysis). Instead, the PSS is derived directly from the orthographic input sentence, by rules which do not refer to a sentence's syntactic structure. Consequently, the resulting PSS can only approximate the theoretical PSS, since not all relevant syntactic information is available for the prosodic analysis.

Some types of syntactic information, however, are indispensable for correct prosodic sentence analysis. For example, the main verb (or verb group) in a sentence must be identified, since this word (group) establishes a separate Phi domain. In Dutch, this Phi domain may separate the subject and object arguments of the verb. Likewise, subordinate sentences must be identified, because they usually establish separate Int domains. In order to establish prosodic domains, however, an exhaustive syntactic analysis appears to be superfluous: it is not necessary to determine the structural relations between the words in a sentence. The necessary information can also be derived from the syntactic word class.

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 the corresponding syntactic word labeling.

3.1. SYNTACTIC WORD LABELING

The distinction between FWs and content words (hence CWs) plays a crucial role in the demarcation of prosodic domains, as explained in section 2 above. Hence, words in the input sentence must be identified as either FW or CW. This is achieved by means of a small lexicon, containing most of the Dutch FWs (544 entries). In addition, this lexicon contains the syntactic classification of each FW. If an input word is found in the lexicon, then its 'prosodic label' is set to FW, and the syntactic label(s) is (are) copied from the lexicon. Words which are not found in this lexicon are given the prosodic label CW. Subsequent syntactic labeling concentrates on these CWs.

For these CWs, syntactic labels are generated on the basis of formal properties of the word string. Affixes trigger the generation of certain syntactic word labels, as illustrated by the following rules (the symbol "|" indicates ambiguity; ":" indicates a sub-classification):

(4a) Undef → (VERB:INFL) / ***dt**
(4b) Undef → (NOUN:PLUR)|(VERB:(INFL|INF)) / ***en**
(4c) Undef → ADV / ***(lijk,ig,zaam)**

In addition, orthographic conventions guide the syntactic labeling: a word containing a hyphen is a compound noun; strings containing digits are labeled as numeral; words starting with a capital (not sentence-initial) are proper names rather than verbs, etc. In the future, the syntactic label(s) for each word will be provided by a separate morphological parser [2, 7], rather than by the above rules of thumb.

Subsequently, words with multiple syntactic labels (either FWs, with labels from the lexicon, or CWs, with labels generated by rule) must be disambiguated. This is done on the basis of restrictions on word order: disambiguation can only be achieved when taking account of the context. The two rules below illustrate this type of 'filtering'. Rule (5) employs a linguistic constraint on word sequences: an inflected verb may not be preceded by an article. Rule (6) employs a statistical constraint: prepositions are usually followed by nouns, rather than by verbs. Such probabilistic observations are based on the analysis of large corpora of newspaper text.

(5) X|VERB → X / ART ___
+ de **bakken** (NOUN)
(6) NOUN|VERB → NOUN / PREP ___
+ vogels broeden in **nesten** (NOUN)
- hij wil zich daar in **mengen** (VERB)

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 the rules of thumb described above.

3.2. PROSODIC ANALYSIS

Since Phi domains never straddle Int domains (see section 2), our algorithm starts with demarcating the latter, higher-order domains. Subsequently, Phi domains are constructed within these Int's. Note that this procedure deviates from the theoretical construction of prosodic domains [13,14].

Several rules demarcate *Int* domains by identifying subordinate clauses. Basically, these rules insert an Int boundary between two adjacent words which cannot both belong to the same clause. In other words, these rules refer implicitly to syntactic conditions on word sequences, and are therefore highly language-specific. Example rules are given in (7-9) below; note that (9) does not apply to English.

(7) 0 → IntBoundary / ___ COMPL
he left his car ## **because** it broke down
(8) 0 → IntBoundary / (VERB:PART) ___ (NOT VERB)
hij is **erin geslaagd ## dit** goed weer te geven
he has succeeded ## to represent this correctly
(9) 0 → IntBoundary / (VERB:INFL) ___ (VERB:INFL)
omdat het geen haast **had ## deed** ik het later
because it no hurry had ## did I it later

In addition, there are also rules which demarcate complex initial constituents (mostly subject NPs). Other rules delimit Int domains on the basis of orthographic interpunction (comma, parentheses, etc.).

For the construction of *Phi* domains, we use the observation that the prosodic head of a Phi is always its rightmost lexical item (see section 2 above). Usually, this head is a CW. Any FWs following this head must belong to a different Phi domain. This yields the following rule, which inserts a Phi boundary between the 'end' of the 'lefthand' Phi (CW head) and the 'start' of the 'righthand' Phi (FW):

(10) 0 → PhiBoundary / CW ___ FW
she **likes # the** red **roses # in** her garden

Of course, this requires special provisions for Phi's with a prosodic head which is not a CW, but rule (7) correctly identifies the majority of Phi domains. Additional rules merge some of the resulting Phi domains (e.g. those containing a single FW) with their lefthand neighbour.

4. PHRASING

At this moment, the PSS is used directly to split the input sentence into separate phrases. Each resulting Int domain becomes manifest as a prosodic break, which is realised in the synthetic speech output as a pause of 250 ms accompanied by an adequate F_0 movement. This procedure, however, results in disfluency in the speech output, which inhibits (rather than facilitates) its correct perception. Apparently, too many breaks

are present in the speech stream, and the phonetic means by which they are realised appear to be too strong.

To improve this situation, two solutions are available. Firstly, the *number* of Int domains in the PSS can be reduced, by means of the Int restructuring mentioned above. In other words, two adjacent Int's are collapsed into a single one (by rules deleting the intermediate Int boundary). Again, this restructuring should be guided by theoretical observations: obligatory Int boundaries (e.g. those based on orthographic interpunction) must be maintained; shorter Int domains are more prone to collapsing; restructuring depends (to a certain extent) on the syntactic function of the Int domains.

Secondly, the *phonetic realisation* of Int boundaries in the output speech could be differentiated, depending on the length and function of the corresponding Int domains. Relatively 'weak' Int boundaries may be realised by phonetic means which are perceptually less salient, e.g. prepausal lengthening rather than an F_0 declination reset.

Although both solutions are currently being investigated, no results are available at this time.

5. ACCENTUATION

5.1. ACCENTUATION: THEORETICAL

In theoretical accounts of sentence accentuation in Dutch, three factors are assumed to be relevant [5,1]. First, a constituent is not accentuated if it contains information which is already "given" ("known, old") to the listener. Only constituents introducing "new" information can be accentuated [11,20], as demonstrated in (11):

(11.1) (he came) (by *CAR*)
(11.2) (his *car*) (was BLUE)

Apparently, listeners rely on this distinction for their understanding of the spoken utterance [16].

Second, the thematic relations between prosodic constituents play an important role [5,1]. As an example, consider the accentuation of the main verb (or verbal group). Even if it conveys new information, this word (group) is usually not accentuated, (12a), unless a non-argument constituent intervenes between the verb and argument (12b).

(12a) ik heb (het gazon met de BLOEMEN) (gemaaid)
I have (the lawn with the flowers) (mowed)
(12b) ik heb (het GAZON) (met de MACHINE) (GEMAAID)
I have (the lawn) (with the machine) (mowed)

Third, rhythmic factors influence accentuation. The occurrence of two adjacent accents (accentuated words) is avoided [10]. In such cases, one of the accents is 'moved' to a different word, as in (13b):

(13a) hij heeft (de hele NACHT) (GELEZEN)
(13b) hij heeft (de HELE nacht) (GELEZEN)

5.2. ACCENTUATION: ALGORITHM

The accentuation component of our algorithm attempts to imitate the theoretical account of sentence accentuation (see also [9]). Firstly, the following categories of words are accentuated: (a) CWs, (b) idiosyncratic words (e.g. **nooit** "never"). Two categories are excluded from this accentuation, viz. (c) verbs and (d) semantically 'empty' words (e.g. **gulden** "guilders"). Subsequently, three types of rules adjust the resulting accentuation:

(14a) rhythmic deaccentuation
(14b) deaccentuation of words conveying given information
(14c) accentuation of verbal constituents in certain contexts

The first rule de-accentuates the middle one of three adjacent accentuated CWs. Both the PSS and the syntactic word labeling are taken into account: all three words must belong to a single Phi domain, and they must belong to certain syntactic categories (e.g. adverb + adjective + noun, quantifier + X + noun). Note that these conditions are stricter than one would expect from (13), in order to avoid incorrect over-applications.

The second type of rule de-accentuates words which can safely be assumed to convey given information. Although the scope of our algorithm is limited to a single sentence, it can nevertheless infer whether words in the sentence under analysis have occurred before. Deictic words, such as **dit, deze** ("this"), **dergelijke, zo'n, zulke** ("such"), imply that the following term conveys information which has already been introduced. The second rule de-accentuates any such words following this cue word in the same Phi domain, as indicated by (15):

(15.1) (he came) (by CAR)
(15.2) (he has BORROWED) (**this vehicle**) (from a FRIEND)

Thirdly, the verb (group) must be accentuated under certain conditions, as mentioned above (see section 5.1.). In order to imitate the effect of thematic relations, for example, a verb (group) is accentuated if it is adjacent to an adverbial phrase, as in (12b) and (16):

(16) ik heb (het GAZON) (VANDAAG) (GEMAAID)
I have (the lawn) (today) (mowed)

In these cases, the adverbial phrase establishes a non-argument constituent, which intervenes between the verb and its argument. Likewise, a verb (group) is accentuated if its argument is un-accentuated; this may happen if the argument contains given information (as in (15.2)), or if the argument is a pronoun.

6. CONCLUSIONS

Many factors which influence human speech communication, are beyond the scope of an algorithm like the one described here. Speech utterances can be realised with a large number of variations, especially with respect to their prosodic phenomena. Human speakers do not make a random selection from all these variants. Instead, their choice for one set of related prosodic phenomena is usually highly adequate, given such factors as the content of the utterance, the context in which it is spoken, and the relations between speaker, listener, content and context. The automatic prosodic sentence analysis and subsequent generation of prosodic markers, as described in this paper, fail to imitate this behaviour of human speakers. This situation will probably remain so for some time, since computer algorithms are not yet able to equal humans in their knowledge of the outside world.

REFERENCES

[1] 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 Rijksuniversiteit Leiden.

[2] BAART, J.L.G., and J.S. HEEMSKERK (1988) The problem of ambiguity in morphological analysis for a Dutch text-to-speech system. In: *Proceedings SPEECH '88 (7th FASE Symposium), Edinburgh 1988.* 3:959-65.

[3] COLLIER, R., and H. 't HART (1975) The role of intonation in speech perception. In: A. Cohen and S.G. Nooteboom (eds.) *Structure and Process in Speech Perception.* Berlin, Heidelberg, New York: Springer. 107-21.

[4] GEE, J.P., and F. GROSJEAN (1983) Performance structures: a psycholinguistic and linguistic appraisal, *Cognitive Psychology* 15:411-58.

[5] GUSSENHOVEN, C. (1984) *On the Grammar and Semantics of Sentence Accents.* Dordrecht, Cinnaminson NJ: Foris. Publications in Language Sciences; 16.

[6] 't HART, J., and R. COLLIER (1975) Integrating different levels of intonation analysis, *J. Phonetics* 3:235-55.

[7] HEEMSKERK, J. (1989) Morphological parsing and lexical morphology. To appear in: H. Bennis and A. van Kemenade (eds.) *Linguistics in the Netherlands 1989.* Dordrecht: Foris.

[8] KAGER, R., and H. QUENé (1987) Deriving prosodic sentence structure without exhaustive syntactic analysis. In: J. Laver and M.A. Jack (eds.) *Proceedings European Conference on Speech Technology, Edinburgh 1987*. Edinburgh: CEP Consultants. 1:243-46.

[9] KAGER, R., and H. QUENé (1989) A sentence accentuation algorithm for a Dutch text-to-speech conversion system. To appear in: H. Bennis and A. van Kemenade (eds.) *Linguistics in the Netherlands 1989*. Dordrecht: Foris.

[10] KAGER, R., and E. VISCH (1988) Metrical constituency and rhythmic adjustment, *Phonology* 5:21-71.

[11] KRUYT, J.G. (1985) *Accents from speakers to listeners: an experimental study of the production and perception of accent patterns in Dutch*. dissertation Rijksuniversiteit Leiden.

[12] te LINDERT, E., C.J. DOEDENS, and H. van LEEUWEN (1989) *Spraakmaker-1*. [Utrecht: Stichting Spraaktechnologie]. SPIN-ASSP Report; 11.

[13] NESPOR, M., and I. VOGEL (1982) Prosodic domains of external sandhi rules. In: H. van der Hulst and N. Smith (eds.) *The Structure of Phonological Representations I*. Dordrecht: Foris. 225-255.

[14] NESPOR, M., and I. VOGEL (1986) *Prosodic Phonology*. Dordrecht: Foris. Studies in Generative Grammar; 28.

[15] NOOTEBOOM, S.G. (1985) A functional view of prosodic timing in speech. In: J.A. Michon (ed.) *Time, mind, and behavior*. Berlin: Springer. 242-52.

[16] NOOTEBOOM, S.G., and J.G. KRUYT (1987) Accents, focus distribution, and the perceived distribution of given and new information: an experiment, *J. Acoust. Soc. Am.* 82:1512-24.

[17] SCHARPFF, P. (1988) personal communication.

[18] SELKIRK, E.O. (1984) *Phonology and Syntax: the relation between sound and structure*. Cambridge MA, London: The M.I.T. Press.

[19] SELKIRK, E.O. (1986) On derived domains in sentence prosody, *Phonology Yearbook* 3:371-405. (C.J. Ewen and J.M. Anderson, eds.).

[20] TERKEN, J.M.B. (1985) *Use and function of intonation: some experiments*. dissertation Rijksuniversiteit Leiden.

[21] WIJK, C. van, and G. KEMPEN (1985) From Sentence Structure to Intonation Contour. In: B.S. Müller (ed.) *Sprachsynthese: zur Synthese von natürlich gesprochener Sprache aus Texten und Konzepten*. Hildesheim: Georg Olms. 157-82.