A RHYTHM-BASED PROSODIC PARSER FOR TEXT-TO-SPEECH SYSTEMS IN FRENCH

Christel SORIN, Danièle LARREUR and Régine LLORCA

Centre National d’Études des Télécommunications 22301 Lannion FRANCE

Abstract

The prosody is one of the main factors deciding the quality of text-to-speech synthesis systems. We present here a system allowing for a prosodic parsing and an automatic prediction of a French prosody which makes no use of syntactic analysis. The system was derived from studies on the prosody used in commercial announcements. In the first step, a sentence is divided into Prosodic Groups (PG’s) which consist of lexical words located between two grammatical words. In the second step, the length and relative location of PG’s determine the insertion of pauses and the specific prosodic categories attributed to each PG. Finally, simple right-to-left derivation rules furnish the prosodic category of each word inside the PG. Predefined F0 and duration rules are then applied depending on the prosodic category attributed to each item.

Introduction

The automatic generation of prosody in text-to-speech system consists into two phases:

Phase 1: definition of prosodic rules allowing to automatically derive F0 and duration contours from prosodic markers (manually) introduced in the text.

Phase 2: definition of parsing rules allowing to predict the location of the prosodic markers automatically.

Existing text-to-speech systems for French include different sets of prosodic rules (see for example, Emerard, 1977, for the CNET synthesis system, O’Shaughnessy, 1984 and Bialy, 1986, for the INRS system, Lienard et al, 1977, for the LMMSI system and Carlson and al, 1982, for the KTH system). These rules were mainly defined by studying F0 contours of read sentences. Another prosodic speaking-style is that used by radio or TV speakers for news or commercial announcements. This “speaking-style” largely uses lexical emphasis and aims to be maximally intelligible and convincing. It could therefore be well adapted to speech synthesis systems towards counter-balancing the negative effects of the segmental defaults of synthesised speech.

In the first part of this paper, we present a new set of prosodic rules trying to mimick French “commercial” prosody. In the second part, the prosodic parser will be described that allows to generate, in the CNET’s synthesis system, both types of prosody the “reading” prosody and the “commercial” prosody.

1- Rules for “commercial” prosody generation in French

The rules system consists into 3 modules:

- a “duration” module
- a “macroprosody” module
- a “microprosody” module.

1/ Duration rules

Two different sets of duration rules were defined. The first one is intimately related to a diphone-based synthesis system. The duration rules aims to complete the duration effects already captured inside the stored diphones by durational modifications which appear inside a sentence. Established 11 rules include the lengthening of the last word-syllable before a main prosodic boundary, the shortening of consonant clusters inside a word, the shortening of middle syllables inside long multisyllabic words, a special treatment for monosyllabic lexical words etc... These rules use the informations provided by the intonation markers which will be described in the following paragraph.

However these rules only modify the intrinsic segmental duration of the stored diphones. Therefore, the criteria used for choosing the diphones (both the environment from which they were extracted and the segmentation criteria) still strongly influence the segmental durations of resulting synthesised sentences.

A second set of rules was developed so that the duration module would be independent from the type of synthesis system (formant or diphone-based). This predictive model of segmental duration (Bartkova et Sorin, 1985) was tested on three corpora : the mean differences between measured and predicted segmental durations were less than the Just Noticeable Difference (JND) for duration in connected speech (Huggins, 1971).
II. Prosodic parsing of a sentence in French

In many text-to-speech synthesis systems, the prosody is derived from words or less complex syntactic analysis of the sentence. However, for French, Choppin and Styczyński (1975) proposed an automatic generation of prosody that avoids the need for a syntactic analysis of the text. Some recent studies (Venk and Hollind, 1982, Dickens and Martin, 1984) suggest that prosodic boundaries could strongly influence the prosodic structure of the text we studied. We observed a strong tendency for segments between pauses or prosodic juncture to have the same number of syllables (generally inferior to 7 syllables).

In these contexts and for practical reasons (i.e. to avoid the use of an heavy syntactic parser), we developed a prosodic parser that maximally uses (beside the punctuation) the presence of short grammatical words inside the sentence. These words have, in fact, main characteristics:

- they are indicators of some syntactic structure
- they present frequently a relatively stable low F0 contour, that acts as a tremping before the higher initial pitch of the following lexical word.

A lexicon of 120 grammatical words was built. The words belonging to this lexicon are marked 6 (they are marked 6) and another group that allows to detect the presence of a verb (they are marked 4) and another group that allows to detect the presence of a verb.

The prosodic parsing of the sentence is done in the following way:

1. Detection of the word marked 6, 6 or 6 which is preceded by a non-verb and before punctuation signs (like *, )... etc.).
2. Sequences of 6 having received the category IV.

The sentence is then parsed into segments between brackets. These segments are derived from "Prosodic Groups" (PG) in the following:

A second module attributes to each PG a specific category which will define the location of the pauses and the main prosodic boundaries. Here, the basic idea was to introduce pauses after long PG in order to simulate breathing pauses, and to use prosodic juncture whenever it was preferable to introduce (in the synthesized sentence) rather larger number of pauses than a realistic number of pauses (as in natural spontaneous speech).

The location of these pauses should be, of course, prosodically plausible.

The final step of the processing consists of deriving the prosodic markers from the categories attributed to each PG. This task is achieved in two different ways for the "reading" prosody in one hand and for the commercial prosody in the other hand. In the first case, a simple correspondence-table associates each category to one of the prosodic defined groups of words (Emard, 1977). In the second case, some...
right-to-left derivation rules are applied inside each PC: a category is attributed to almost every word in the sentence (some intermediate rules group some monosyllabic word sequences into an unique "prosodic word"). At this level, (which now use 6 categories) a correspondence table associates to each word-category one of the markers which were presented in the first part of this paper (Table III).

### TABLE III

<table>
<thead>
<tr>
<th>Category</th>
<th>Prosodic Marker</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0</td>
</tr>
<tr>
<td>II</td>
<td>4*</td>
</tr>
<tr>
<td>IV</td>
<td>1- or 5- (monosyllabic)</td>
</tr>
<tr>
<td>V</td>
<td>4*</td>
</tr>
<tr>
<td>VI</td>
<td>3</td>
</tr>
<tr>
<td>$\phi$, $\phi^a$, or $\phi^b$ word</td>
<td>6</td>
</tr>
<tr>
<td>unique</td>
<td>6</td>
</tr>
<tr>
<td>two</td>
<td>6 and 6-</td>
</tr>
<tr>
<td>short pause “p”</td>
<td>8</td>
</tr>
<tr>
<td>long pause “p”</td>
<td>7</td>
</tr>
</tbody>
</table>

Table IV gives some examples of the results both for the PG categorization and for the allocation on prosodic markers for the "commercial" prosody.

### Conclusion

The entire prosodic module was tested on a large body of TELEX messages. Special items like surnames, acronyms, numbers, abbreviations, were treated beforehand by a text-preprocessing module. The results were judged to be satisfactory enough to implement this module into a text-to-speech system for reading electronic mail. Some defaults of this module indicate the limits of a "syntax-independent" prosodic parser: in some cases, rhythmic constraints must be subordinated to syntactic structure, which cannot be detected without a profound syntactic analysis. This is the case, in particular, for verbs or verbal forms, as illustrated in Table IV ("mis en place" must be considered as a PG because it is derived from the verbal form "mettre en place"). Corresponding prosodic improvements could then be reached only in using, at least, a large lexicon of verbal forms or a fine syntactic (and maybe) semantic analysis which remains to be done.

### Acknowledgements

The authors wish to thank S. Maeda for his help in writing the English version of this paper.

### References