A RYTHM-BASED PROSODIC PARSER FOR TEXT-to-SPEECH SYSTEMS IN FRENCH

Christel SORIN, Danièle LARREUR and Régine LLORCA

Centre National d'Etudes 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 Fo 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 Fo 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 sytem, O'Shaughnessy, 1984 and Bailly, 1986, for the INRS system, Lienard et al, 1977, for the LIMSI system and Carlson and al, 1982, for the KTH system). These rules were mainly defined by studying Fo 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 system towards counterbalancing the negative effects of the segmental defaults of synthetised 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 paser will be described that allows to generate, in the CNET's synthesis system, both types of prosody the "reading" prosody and the "commercial" prosody.

I- 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 diphones-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-syllabe before a main prosodic boundary, the shorthening of consonant clusters inside a word, the shortening of middle syllables inside long plurisyllabic 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.

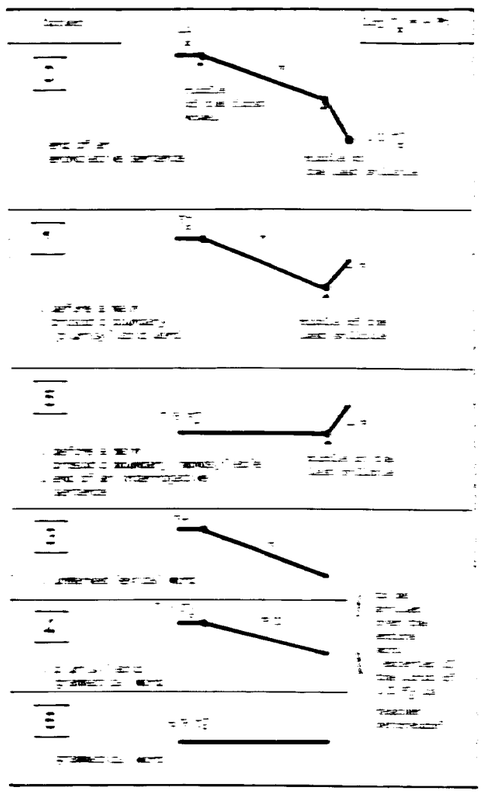
However these rules only modify the intrinsic segmental duration of the stored diphones. Therefore, the criteria used for choosing the diphones (both the environment from which they were extracted and the segmentation criteria) still strongly influence the segmental durations of resulting synthesised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

A second set of rules was developped 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).

2 Macroprosodic rules :

Macroprosodic rules were based on the study of the prosody used in commercial announcements. The basic corpus contained 100 sentences recorded by a professional female speaker for a commercial purpose (informations on a new product). The observed Fo contours were well described by 4 typical Fo contours for which the Fo evolution was formalised. Table I presents these 4 schematized Fo contours that are associated to 4 prosodic markers. They can apply either on a word or on a sequence of words located at the left of the prosodic marker inserted into the sentence. These contours are mainly defined by their Fo initial value (Fo^I), the slope P, the final Fo value for the sentence final Fo contour. Two supplementary markers - and = can be associated whith any marker. They allow to increase or decrease the Fo initial value by 1 or 2 Fo^I steps. For long words or words-sequences involved in an unique Fo contour, some rules modify slightly the above presented contours (for example, the P-slope is divided by a factor 2, the Fo^I value is maintained over the 2 first syllables etc... .

TABLE I



at this level, the pauses are indicated by 2 supplementary markers " for a long 300 ms pause, ' for a short 100 ms pause that can be associated with every other Fo marker.

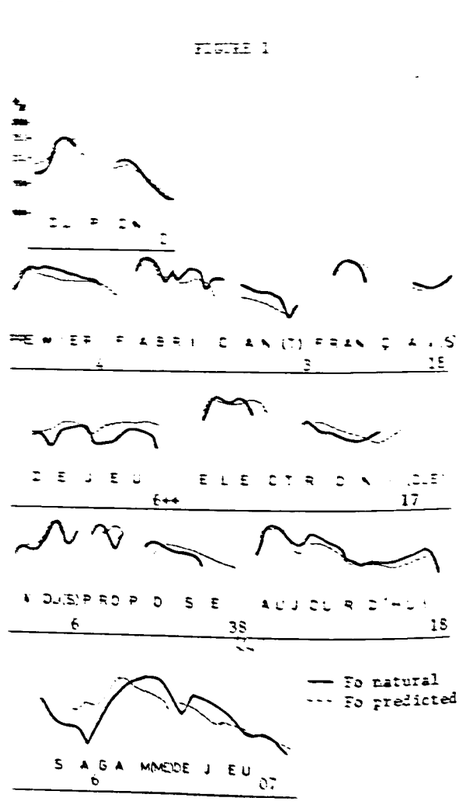
3 Micromprosodic rules :

This module contains three set of rules which are successively applied :

- micromprosodic rules for vowels in an unvoiced context : smoothering of the vowel Fo contour when preceeded or followed by an unvoiced segment),
- micromprosodic rules for voiced consonants : a dip is introduced in the macroprosodic Fo contour at the place of the consonant,
- "micrifluctuation" rules : to avoid the presence of long sequences having the same Fo value (for example long vowels) some fluctuations are introduced on the flat Fo contours (their magnitude is less than 11 Hz).

This set of rules allows for a good prediction of the observed Fo contours. As an example, Figure 1 displays the Fo contours that were obtained after manual assignment of prosodic markers, in comparison with the original Fo contours of the sentence.

FIGURE 1



II. Prosodic parsing of a sentence in French

In many text-to-speech synthesis systems, the prosody is derived from more or less complex syntactic analysis of the sentence. However, for French, Choppy and al (1975) proposed an automatic generation of prosody that avoids the need of a syntactic analysis of the text. Some recent studies (Wenk and Wioland (1982), Dell (1984) and Martin (1986)) suggest that rythmical constraints could strongly influence the prosodic structure of the sentence. In the corpus 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 context and for practical reasons (i.e. to avoid the use of an heavy syntactic parser), we developped a prosodic parser that maximally uses (beside the ponctuation) the presence of short grammatical words inside the sentence. These words have, in fact, 2 main characteristics :

- they are indicators of some syntactic structure
- they present frequently a relatively stable low Fo contour, that acts as a trempling 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 ϕ. Among them, a special group contains the grammatical words that, most of the time, introduce a subordinate phrase (they are marked ϕ**) and another group that allows to detect the presence of a verb (they are marked ϕ*).
The prosodic parsing of the sentence is done in the following way :

1/ detection of the word marked ϕ, ϕ* or ϕ**
2/ introduction of brackets (][) before every word ϕ, ϕ* or ϕ** which is preceeded by a non ϕ-word and before ponctuation signs (like ", ( ) :" etc...).

The sentence is then parsed into segments between brackets. These segments will be designated as "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. We hypothesised that it was preferable to introduce (in the synthesised sentence rather larger number of pauses than a realistic number of pauses (as in natural spontaneous speech : such pauses could reduce the mental load of the listener due to the heavier processing of alterated speech (Nusbaum and Pisoni, 1982). However, the location of those pauses should be, of course, prosodically plausible.

4 main categories are attributed to each PG as a function of :

- the number of lexical words inside each PG
- the position of the PG inside the sentence
- in some cases, the number of syllables in the PG and the previously attributed categories of the surrounding PG's.

TABLE II

Examples of prosodic parsing rules

- the sentence-final PG	. receives the category I
- PG followed by a comma	. receives the category IV . is followed by a long pause "P"
- PG containing 3 (or more) lexical words	. receives the category IV . is followed by a long pause "P" . attributes the category IV to the preceeding PG . is preceeded by a short pause "p" (facultative)
- PG followed by a PG containing a ϕ* or ϕ**-word	. receives the category IV . is followed by a short pause "p"
Stylistic rule (specifically observed in commercial announcements) - PG preceeding the sentence final PG	- if the total syllables number of the 2 PG's exceeds 7 syllables : . receives the category IV . is followed by a short pause "p" - if not : . receives the category II . attributes the category IV to the preceeding PG . is preceeded by a short pause "p"
- PG containing an unique lexical word	. receives the category V (if no category was previously attributed)
- PG containing 2 lexical words	. a set of contextual rules attribute or the category V or the category IV and a short pause
- sequences of PG having received the category V	. if the total number of syllables exceeds 7, an eurythmic index is calculated : a short pause is introduced between the PG's which delimit the eurythmic structure. Category IV is attributed to the PG preceeding this pause.
- etc...(essentially Pauses-harmonisation Rules).	

Right-to-left derivation rules inside a PG

V ← VI ← V ← II ← I
V ← VI ← V ← IV
V ← VI ← V

The final step of the processing consists of deriving the prosodic markers from the categories attributed to each group. This task is achieved in 2 different ways for the "reading" prosody in one hand and for the commercial prosody in the other hand. In the first case, a simple correspondance-table associates each category to one of the previously defined prosodic markers (Emerard, 1977). In the second case, some

right-to-left derivation rules are applied inside each PG : 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 correspondance table associates to each word-category one of the markers which were presented in the first part of this paper (Table III).

TABLE III

Category	Prosodic Marker
I	0-
II	4*
IV	1- or 5- (monosyll.)
V	4*
VI	3-
ϕ, ϕ* or ϕ** word . unique . two . sequence	 6 6 and 6- 4-
short pause "p"	8
long pause "P"	7

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, rythmical 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 an 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 may be) 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.

Références

Bailly, G. (1986) : "Multiparametric generation of French prosody from unrestricted text", IEEE-ICASSP, 2419-2422.

Bartkova, K. and Sorin, C. (1985) : "Predictive model of segmental durations in French", J. Acous. Soc. Am., 77, suppl. 1, S54 (to appear in speech Comm.).

Carlson, R., Granström, B. and Hunnicutt, S. (1982) : "A multi-language text-to-speech module", Proc. IEEE-ICASSP 82, 1604-1507.

Choppy, C., Lienard, J.S., and Teil, D. (1975) : "Un algorithme de prosodie automatique sans analyse syntaxique", Proc. 6th JEP/GALF, 387-395.

Dell, F. (1984) : "L'accentuation dans les phrases en français", in "Formes sonores du langage", ed. by Dell, Hirst and Vergnaud, Hermann, Paris, 65-122.

Emerard, F. (1977) : "Synthèse par diphones et traitement de la prosodie", Thèse 3° cycle, Univ. of Grenoble.

Huggins, A.W.F. (1971) : "Just noticeable differences for segment duration in natural speech", J. Acous. Soc. Am., 51, 1270-1278.

Lienard, J.S., Teil, D., Choppy, C. and Renard, G. (1977) : "Diphone synthesis of French : Vocal response unit and automatic prosody from text", Proc. IEEE-ICASSP 77, 560-563.

Martin, P. (1986) : "Structure prosodique et structure rythmique pour la synthèse", Proc. 15 JEP/ GALF, 89-91.

Nusbaum, H.C. and Pisoni, D.B. (1982) : "Perceptual and cognitive constraints in the use of voice response systems", Research on Speech Perception, Progress Report 8, Indiana Univ., 203-216.

O'Shaughnessy, D. (1984) : "Design of a real-time French text-to-speech system", Speech Comm., 3, 233-243.

Wenk, B. and Wioland, F. : "Is French really syllable timed", J. of Phonetics, 10, 193-216.

TABLE IV : **Examples of Prosodic Parsing and Allocation of Prosodic Markers**
(sentences presenting no ponctuation sign)

P = long pause
p = short pause

ϕ words	ϕ*	ϕ**	ϕ*	ϕ	ϕ	ϕ	
PG categories	IV	IV	V	V	IV	I	
Intra-PG categories			VI			V	II
Prosodic markers	[Demain] [vous qui voyagez] [vous pourrez gagner] [du temps] [en utilisant] [les voyages vite Air]						
	1S- 6	6- 1S-	5 3	4* 5	3- 5	1S- 6	4* 3- 07-
Prosodic Parsing and PG categories	[Les trois malfaiteurs] [et le complice] [qui les attendait] [au volant] [d'une voiture]						
	V	IV_p	V	V	IV_p		
						mis][en place]	
	[ont réussi] [à échapper] [aux policiers] [en dépit] [de l'important dispositif					IV_p V	
	V	IV_p	V	IV_p		policier][en place]	
						IV_p	
	[dans toute la région] [en emportant] [un butin] [dont le montant] [n'a pas été révélé].						
		IV_p	V	IV_p	IV_p	I_p	