Prosody modeling with a dynamic lexicon of intonative forms: Application for text-to-speech synthesis

Véronique Aubergé
Institut de la Communication Parlée, INPG/ENSERG - Université Stendhal, BP 25X, 38040 Grenoble, France
Email : auberge@icp.grenet.fr

ABSTRACT

We propose here a methodology and tools for the semi-automatic constitution of a intonative generation module. .

The first stage of the work was the analysis of a corpus recorded by a reference speaker and based on a set of linguistic presuppositions. These presuppositions are based on the concept of some structural ***rendez-vous*** *between the different levels of text on one part and the prosody. on the other part The processing of the data corpus was organized in a top-down hierarchy: sentences, clauses, groups and lexical units. The minimal symbolic unit is the syllable. For every level in the hierarchy, several initial classes of Fo contours are defined, each initially described by a maximal set of linguistic parameters. The validity of each class is first verified. Then the unification of the classes is systematically tested, using minimal pairs oppositions on the linguistic parameters. For every final class an average-contour is computed, which is a global form for this class. The result is a hierarchically structured* ***dynamic contour lexicon*** *of* ***global intonative forms*** *for which every representative is associated with a minimal set of distinctive attributes.*

Generation of prosody then consists in the calculation of prosodic patterns by the top-down cumulative superpositions of contours taken from the lexicon. An application is the automatic generation of prosody in a text-to-speech synthesis system, which must be adapted to a given application.

INTRODUCTION

The main difficulty that text-to-speech (TtS) synthesis has to deal with, is the generation of a well formed intonation on the sole basis of textual information and a few eventual extra-linguistic parameters concerning the speaker(s) and communication situation. Intonation is, nevertheless, a verbal language phenomenon (Bolinger, 1989). The transition from text –i.e., from written language– to speech is rather artificial or, at least, somehow special. Certain theories have suggested congruency relationships (that is, partial identity projections) between syntactic and intonative structures. We have adopted, in this study, a less restrictive perspective: intonation per se carries specific information and refers in fact to a particular performance structure of spoken language (Monnin & Grosjean, in press). The great coherency of the language structures implies that other linguistic structures are covered by intonation as well, in a somehow redundant fashion, especially in read-out text. This suggests the existence of *rendez-vous* nodes, from which sub-structures could be isolated. Specifically, this refers to the isolation of the sub-structures' global form, without expecting, however, to interpret its logic.

The aim was to propose a methodology with which a module could reproduce a specific utterance from a corpus (of a single speaker in a defined communication situation). This phonetic modeling corresponds to a lexicon of global intonative forms. It is structured according to a hierarchy of *rendez-vous* node levels. This lexicon, which is specific to the speaker and situation, links a symbolic input with a physical output, without explicitly involving the phonological level, although it is implicit in the structure of the lexicon.

CORPUS DESIGN

Prosodic coding

The prosodic parameters selected for this study are the fundamental frequency F0 and duration, and the minimal suprasegmental unit for both is the syllable (Campbell, 1992) even if this choice is not necessarily definitive (Barbosa & Bailly, 1993). The relative importance of F0 and duration in a given utterance context is subject to the choice of the speaker (Caelen, 1992).

For the segmental level, F0 was coded with three values (beginning, maximum excursion, end) for each syllabic nucleus. The duration was coded in two successive stages. For the segmental level, start and end points were stored for each vowel. For the supra-syllabic level, syllabic durations are measured on the stored syllabic F0 contours. Then, a measure termed *prosodic syllable length* is calculated for every syllable. Its intent is to quantify that part of the syllable duration which is not ascribable to inherent segmental phenomena, and it is calculated by subtracting the duration of a reference syllable from that of the syllable in the natural corpus.

Constraints: minimal pairs

The first step of the study presented here is the constitution of a corpus based on strict constraints of *minimal pair oppositions* which are supposed to give prominence to what we refer to as *rendez-vous* between the intonation and the linguistic structure nodes. The linguistic units are hierarchically organized: the highest level treated is the *sentence*, while the lower levels consist of the *clause*, the *group*, and *subgroup* (experimentally, it was confirmed that the subgroups exist inside the groups of 5 or more syllables long).

The factorial design of the corpus is based on the combination of minimal oppositions within attributes at each syntactic level. The basic opposition at the sentence level is one of sentence modality: declarative interrogative, or imperative. These modes are subdivided in fairly classic ways, declarative into positive and negative, and interrogative into direct, introduced, inverted, and elliptic. Sentence length in syllables was also treated, via sentences from 3 to 21 syllables long (see examples in Table 1).

At the clause level the oppositions concern syntactic dependency or embedding, linear position, and clause length. If either a clause is embedded or it is the matrix clause in which another clause is embedded, it is classed as *linked*; if not, it is *unlinked.* Unlinked clauses may be isolated, coordinated, or juxtaposed. Dependent clauses may be verb dependent or noun dependent (see examples in Table 1). The clause may be in initial, medial, or final position in the sentence, and length in syllables runs from 3 to 9.

The length of the groups varies from 2 to 15 syllables. Different values are defined for the nature of the group: nominal, verbal, adjectival, adverbial, and "grammatical word" group. The nominal group (NG) has been more most closely studied. When the NG is over 4 syllables long, it is arbitrarily decomposed into sub-groups. For small NGs (i.e., 4 syllables or fewer), the group is the terminal level. The functions of NGs are subject, object and complements. The absolute position inside the clause is an attribute of the group. A relative position attribute is used, in particular for the NG in front of the verbal group. During the analysis phase, this attribute was revealed to be redundant with the subject attribute.

A sub-group is a constituent of NG more than 4 syllables long. It can vary between 2 and 12 syllables. If the higher group is a syntactically simple one, the next level of constituents is the word. This type of sub-group is characterized by the categorial values of the word (grammatical word, adverb, adjective, noun, auxiliary, conjugated or composed verb) and by the relative position of the adjective in front of the noun. If the higher group is complex, the decomposition depends on the nature of the group structure: for an enumeration structure, the sub-groups each correspond to an enumerated constituent, while in case of a dependency structure the dependent constituent is treated like a group. Also, this last case, where the sub-group is "promoted" to group status, is the only case where the sub-group is not also a terminal level.

The morpho-syntactic values represented in the corpus are equivalent to the output of the automatic text analysis performed Stefanini et al.(1992) at CRISTAL (Grenoble).

Table 1. *Selected examples of the successive levels*

attributes in the sentence level		**examples**
interrogative	direct	Tu viens ?
	inverted	Viens-tu ?
attributes in the clause level		**examples**
linked	subord. verb->clause	Je verrai *si les enfants jouent*
attributes in the group level		**examples**
length	3 syllables	*Les enfants* jouaient sur la chaussée
pos. in the clause	initial	*Les enfants* jouaient sur la chaussée
function (NG)	object	Je vois *les enfants.*
	verb dependent	Les enfants jouaient *sur la chaussée*
attrib. in the sub-group level		**examples**
pos. in the group	before noun	Les *petits* enfants jouaient sur la chaussée.

METHODOLOGY OF ANALYSIS

The database, once recorded, was constituted according to the design described above, associating the phonetic labels, the syllable boundaries, the segmental and syllabic F0, the segmental and syllabic duration codes, and the linguistic attribute values.

The main hypothesis underlying this methodology is that each linguistic level (i.e., sentence, clause, group, and sub-group) corresponds to a *global intonation unit*, which can be performed by varying intonation forms (extracted from contours). At the lowest (terminal) linguistic level (sub-group, or group, as appropriate), the contours of the unit , which is also a terminal intonation unit, are defined directly for each syllable by the *syllabic F0* and the *prosodic syllable lengths* . At the higher linguistic levels, the contours of the unit, which is a non-terminal intonation unit, are defined by length attribute (number of syllables) in all cases, but by the *syllabic F0* exclusively for the first and last syllable of the unit considered (i.e., sentence, clause, or group). The intonation contours will thus be modeled by a simple declination line for the non-terminal units and by a (short) string of values for the terminal units.

These different intonation contours in the corpus were automatically segmented for each level, and were exhaustively and hierarchically grouped at all levels according to all the attribute values applicable to that contour at that level.

Further analysis consisted of verifying the homogeneity of intonation contours according to the different combinations of linguistic attributes, and this was done at every level. For the non-terminal levels, which are declination lines, statistical criteria were used, but in the absence of objective criteria for judging similarity of the terminal contours, where values are defined for each syllable, we used visual comparison.

Thus, at each linguistic level, classes of intonation contours are defined, each indexed by a set of attribute values. A representative contour, termed the mean contour (henceforth MC), is calculated for every class.

A lexicon, hierarchically following the linguistic levels, is then constituted with all the MCs, each indexed by the appropriate attribute values.

GENERATION OF INTONATION IN TtS

The text input to the synthesizer is processed sentence by sentence. The automatic text-to-phonetics and morpho-syntactic analysis yields the linguistic attributes. At the suprasegmental level, generating the intonation for an input sentence consists of the hierarchical calculation of a syllabic F0 pattern from the first, non-terminal level (the sentence) down to the terminal level (the group or the sub-group, as appropriate) and finally, of a prosodic syllable length pattern as well, though only at the terminal level. The declination line of the sentence is fixed first, and then this pattern becomes the input pattern of the clause level calculation module. The average-contours of the sentence's component clauses are juxtaposed to define a "local" declination line for the sentence, and this local pattern is warped to the input pattern. The resulting pattern becomes the input pattern for the group level calculation module. When groups are divided into subgroups, the group's MC is warped the same as those of the higher levels, and processing continues to the terminal subgroups. Since the average-contours of the subgroups and terminal groups are not simply declination lines, one is calculated in order to guide the warping process for the terminal group or subgroup. Finally, the prosodic syllable lengths are retrieved and associated to the F0 pattern. The output of the suprasegmental calculation of the intonation is a pair of functions specified in discrete syllabic steps, one for F0 and one for prosodic syllable lengths.

The last step is the "segmental" calculation of the intonation pattern. We have seen that F0 is coded with three values for each vowel. These values control the synthesis at the beginning, the middle and the end of the vowel. The acoustic curve of F0 is calculated by a Spline function interpolation between these points, and the curves for consonants or clusters are interpolated between vowels with the same Spline function.

The durations of the syllables are modified by the intrinsic and co-intrinsic values already used during the analysis of the corpus.

CONCLUSION

The hypothesized *rendez-vous* between prosodic strategy and the corpus structure design might have been invalid, either generally or for our speaker. In fact, the results show that the *rendez-vous* was quite clearly made and consequently the resulting synthesized intonation is high quality. That can be explained partly because of the specific situation chosen, a set of sentences read in isolation without any coherence outside the sentence level.

One unsatisfying point in this study is the duration processing. Presently, it is debatable whether it is the syllable or the group inter P-Center (Barbosa & Bailly, 1993) that is the appropriate unit to be chosen.

An application of this method of corpus constitution and analysis should be considered for other applications. Concerning a given language, an extension of this work could include building a database of MCs representative of different prosodic strategies (man-machine dialogue, multi-media bureautics, or remote control...).

The explicitation of the phonological level underlying the lexicon is one way to generalize such model (Hirst & Di Cristo, 1992). Another way is the systematic unification of the classes of MC, or the learning thrue a stochastic model such a neural network (Traber, 1992).

REFERENCES

V. Aubergé (1992), "Developing a Structured Lexicon for Synthesis of Prosody", In *Talking machines: theories, models and designs* , ed. by G. Bailly & C. Benoît (North Holland Pubisher), 307-322.

P. Barbosa, G. Bailly (1993), "Generation and evolution of rhythmic patterns of text-to-speech synthesis", *this volume,*

D.L. Bolinger (1989), *Intonation and its uses: Melody in gramar and discourse* Stanford, CA: Stanford University Press.

G. Caelen-Haumont (1993), " Dialogue homme-machine et intelligibilité : analyse des caractéristiques linguistiques et prosodiques des discours de lecture,*Séminaire Prosodie, GDR-PRC,* 113-128.

W. N. Campbell (1992), "Syllable-based segmental duration.", In *Talking machines: theories, models and designs* , ed. by G. Bailly & C. Benoît (North Holland Publisher), 211-224 .

F. Grosjean & J.Y. Dommergues (1983). "Les structures de performance en psycholinguistique," *L'Année Psychologique,* 83, 513-536.

D. Hirst & A. Di Cristo (1992), "Niveau de représentation et étiquetages prosodiques", *Séminaire Prosodie, GDR-PRC Communication Homme-Machine,* Aix-en-Provence.

P. Monnin & F. Grosjean (in Press), " Les structures de performances en français : caractérisation et prédiction," *L'Année Psychologique.*

M.H. Stéfanini, A. Berrendonner, G. Lallich & F. Oquendo (1992), "TALISMAN: A multi-agent system governed by linguistic laws for natural language processing", *Proceedings of COLING,* 490-497.

C. Traber (1992), "Fo generation with a database of natural Fo patterns and with a neural network." In *Talking machines* , ed. by G. Bailly & C. Benoît (North Holland Publisher), 287-304.