MULTILINEAR MODEL OF FRENCH PROSODY IMPLEMENTED ON A TEXT-TO-SPEECH SYSTEM

S. Barber, D. J. Hirst**, J. House***, P. Nicolas**, P. Roméas** and B. Waernulf**
(alphabetical order)

**Telia Promotor Infovox AB, Sweden.*
***Laboratoire CNRS URA 261 "Parole & Langage", Université de Provence, France.*
****Department of Phonetics and Linguistics, University College London, England.*

ABSTRACT

This model provides a TTS system with a multilinear hierarchical approach to prosody. Text is processed through two levels of constituents parsing : Intonation Units and Accent Groups. The acoustic realisation level is reached after rhythmic and tone sequence adjustments.

INTRODUCTION

This model for the French language was implemented on the Infovox TTS system. The RULSYS [1] rules format allows a formalism which is quite close to that of generative grammar. The model provides a stepwise derivation from underlying abstract levels of prosody to acoustic realisation level, which makes it quite different to what previous versions of the system did [2]. Most of the representations presented here rely on recent developments of multilinear phonology, which are the general framework of theoretical and experimental works on prosody at the Aix Institute of Phonetics [3,4].

PHONOLOGICAL REPRESENTATION

It is assumed that prosody has an autonomous level of representation, which means that it can to a certain extent be predicted independently from the syntactic and lexical material of the text. General prosodic pattern to be generated for any utterance is known prior to the syntactic analysis of the text. The identification of prosodic constituents is carried out through an algorithm that looks for syntactic markers and labelled parts of speech in the text. This provisional parsing may later be called into question according to rhythmic constraints or to tonal phonotactic constraints.

LEVELS AND CONSTITUENTS

It is assumed that all French utterances consist of a sequence of Intonation Units (IU), the deepest constituent level. A second level is represented by Accent Groups (AG). The extent of AGs is subordinated to the extent of IU : no AG is allowed to spread on both sides of a IU boundary. No AG can exist out of an IU constituent, whereas an IU may (in some extreme cases) not be parsed into AGs.

IUs can be terminal or non-terminal. All sentences have a single terminal IU. The number of non-terminal IUs in a sentence can be 0 and is theoretically unlimited. The assignment of IUs and AGs relies on the assumption that two types of syllables may potentially be assigned stress in French : word initial and word final ones. Assuming that the syllable is the phonological unit that carries stress, the IU must be considered a prosodic constituent made of a sequence of unstressed syllables followed by one stressed syllable (i.e. the IU head). Thus French is described as having right headed major prosodic constituents. IU stress is always on word final syllables, regardless to word class (+ or - OPEN). The system's feature <PSTRESS>, which can be attached to vowels, is used to identify this stressed syllable.

There is no lexical prosody in French: no distinctive stress feature belongs to the definition of lexical units. Yet the lexical unit is the domain of secondary stress in French [5,6], the primary stress being represented by IU heads. Our assumption of secondary stress refers to both lexical initial pitch accent and lexical final lengthening (except IU boundaries). Such a grouping defines what we call Accent Group (AG). Lexical units are identified by the feature <+OPEN> in the system lexicon. AG stress is exclusively assigned to *initial* and *final* syllable of the lexical unit. So at this deep level of representation, any lexical unit boundary syllable is potentially stressable, but no AG head is specified. The system recognises these syllables as their vowels carry the feature <ASTRESS>. At a later stage of processing, deletion rules, lengthening rules, and context sensitive rules provide appropriate tone and/or duration interpretation of <ASTRESS>.

IU-PARSING

IU-parsing consists of finding the end of every IU in the sentence. It is carried out through two successive levels of boundary assignment. The first level of IU boundary assignment deals with no longer deletable boundaries. The second level deals with IU boundaries that may be deleted according to rhythmic constraints. If not deleted, these second level boundaries are interpreted as tones. The way rhythmic constraints apply is explained below.

At the first level, boundaries are exclusively determined by punctuation marks, lexicon labels, or by some markers given as output of the system syntax module, without any rhythmic determination. For example, a comma always generates the same set of four features of the system, which itself is always interpreted as a given tone, a given lengthening, and a 25 frame pause. It is important to point out that this is an exception to the general idea of this model, since all other levels of prosodic categories assignment do not rely on term to term relations between some module output and one prosodic category. All other prosodic categories are subject to later contextual changes and deletions.

Boundaries of terminal IUs are generated at the first level. Punctuation marks ".", "?" and "!" indicate them. Depending on terminal punctuation mark and on the presence of <WH> question feature in the sentence, these boundaries will be interpreted by four possible tones at the realisation rules level. The last vowel in these IU are assigned features <PSTRESS, TERM> which lengthening rules will recognise as to be interpreted as the maximal lengthening (i.e. L5).

As for questions, a distinction is made between : wh-, yes/no, and "A ou B?" questions (consisting of two sections separated by the conjunction "ou"). This distinction relies on text and part of speech labels carried by the words.

Other punctuation marks "," ";" ":") as well as "(" and ")" also generate a first level boundary at their left. Some of them convey features that are responsible for further tone reduction. Lengthening in these cases is L4. Tones are defined in the system by sets of binary features. Lengthening is defined as a percentage of the default durations given in the definition module of the system.

A few connective words are IU initiators in French, and were labelled as such in the lexicon. They always generate a first level boundary at their left. They are a small list of previously <-OPEN> words, mostly conjunctions, now turned to <+OPEN> in the present system. Some of them get <+OPEN> in some restrictive morpho-syntactic contexts. Their label creates a specific feature at the appropriate processing stage.

The end of a relative clause <ENDREL> also generates a first level IU boundary. <ENDREL> can coincide with any punctuation mark, or with the beginning of the main clause verb. The main problem here was to identify this verb.

As to the second level of IU boundaries, syntactic markers initiate parsing but rhythmic balance rules may delete boundaries even though they were assigned at major syntactic constituents ends. Nevertheless, rhythmic rules do not absolutely ignore the boundary depth.

At a first stage of assignment, many IU boundaries are generated, each being assigned a rank which depends on syntax markers, part of speech labels and syllable count. Undeletable IU boundaries (i.e. first level) are assigned rank 0. Then processing is as follows :

1) Assign a rank 1 IU boundary in front of a relative clause marker.

2) Assign a rank 2 IU boundary in front of a prepositional phrase marker.

3) Assign a rank 3 IU boundary after a verb (identified by its PS-label).

4) Assign a rank 4 IU boundary in front of a verb phrase marker.

5) Assign a rank 5 IU boundary in front of any remaining phrase marker.

6) Assign a rank 6 IU boundary after the final syllable of a lexical unit (<+OPEN> word) if there are at least 4 syllables between it and the preceding boundary.

This of course gives far too many boundaries and a selective deletion process is needed. Rank by rank, starting from rank 6, rules delete as many boundaries as possible, namely as long as the number of syllables of the new formed IU is not over 10. The probability for a rank 1 boundary to be deleted is thus lower than the probability for a rank 6 boundary. A short sub-algorithm can be used in order to prevent leaving very short IUs either at the end or at the beginning of a sequence of words bounded by rank 0 boundaries after the deletion process.(See line (2) below).

In the following representation of a sentence, "x" stands for a syllable, a digit between two "x" represents a boundary with its rank, and "D" stands for "deleted". Each line represents a step in the derivation.

(1) xx0xxx1xx0xxx3xx2xxxx5xx2xx1xx0xxx0
(2) xx0xxx1xx0xxx3xx2xxxx5xx2xxDxx0xxx0
(3) xx0xxx1xx0xxx3xx2xxxxDxx2xxDxx0xxx0
(4) xx0xxx1xx0xxxDxx2xxxxDxx2xxDxx0xxx0
(5) xx0xxxDxx0xxxDxx2xxxxDxx2xxDxx0xxx0
(Result) xx0xxxxxx0xxxxx2xxxxxxx2xxx0xxx0

After deletions, remaining boundaries receive a tonal interpretation. At this level, tones are opposed to each other using a set of binary features (terminal or not, rising or falling, expanded range or not, reduced range or not) which are coded in the system. Reduction of range can occur depending on a syllable count threshold. Later on, in AG processing rules, another series of tones, of a different type, is needed. All tones are interpreted as acoustic Fo patterns at the end of all prosodic rules.

Relative clause, interrogative or exclamative contexts impose boundary transformations. As a summary :

- Any /+FoRISE/ boundary tone that is located inside a relative clause is turned to /-FoRISE/. Connected relative clauses are also specifically treated.

- All 3 types of questions require specific tone adjustments or assignments in non-terminal locations, since it is assumed that specific tonal marks of question occur in 3 locations in the sentence, not just at the end. These were called the "secondary" (sentence internal) and the "tertiary" (sentence initial) question marks. First a type-specific transformation of one of the non-terminal IU boundaries is made. Whatever the type of the question is, all non-terminal boundaries that follow the secondary question mark change /+FoRISE/ feature to /-FoRISE/. Eventually, another particularity of all questions is that all boundaries that are located between the secondary question mark and the preceding rank 0 boundary (if any) also reverses this feature polarity. Exclamative sentences are processed as questions in a first step. Then at the end of all parsing algorithms, some transformations are made. The tertiary question mark is a matter of AG stress assignment and is treated thereafter.

AG-PARSING

After IU-parsing is carried out, no IU boundary tone can be deleted any longer. Although AG stresses can only be assigned to initial and final syllables of <+OPEN> words, an AG stress should never be assigned to a syllable that has already been assigned an IU boundary. If the sentence is a yes/no question, an "A ou B?" question or an exclamative sentence with no <WH> word, then an AG stress is assigned to the first syllable of the first word in the IU ending with the secondary question mark. This is the only exception where an AG stress may be assigned on a <-OPEN> word. This initial AG stress will be interpreted by a specific higher tone later on, representing a tertiary question mark. As said earlier, all vowels that belong to a syllable that is assigned an AG stress take the feature <ASTRESS>. The first AG stress in any IU (AG1) is undeletable, except by a non-terminal /+FoRISE/ IU boundary tone located on its right adjacent syllable. If the sentence consists of just one terminal IU, this undeletable AG1 shifts to the end of the next word-final syllable located on a polysyllabic item. The undeletable peculiarity of this AG stress is coded by a feature on the vowel.

IU boundary tones never delete an adjacent AG stress that belongs to the next IU to the right (No cross-IU deletions). Non-terminal /+FoRISE/ boundary tones delete any AG stress on an adjacent syllable to its left. So do /-FoRISE/ boundary tones, terminal or not, except with AG1.

Then comes the AG deletion stage : proceeding from left to right, each AG stress deletes the next deletable AG stress

if it is located on the immediately following syllable. Example :
"*le CHAPEAU du MAIRE de MARSEILLE(...)*" becomes :
"*le CHApeau du MAIRE de MarSEILLE(...)*", in which "peau" is deaccented by AG stress on "cha", and "Mar" is deaccented by non-terminal rising IU boundary tone on "seille".

A couple of somewhat ad hoc but very useful tones were created, namely <ds> (for DownStep) and <rs> (for ReSet). <ds> is assigned under some conditions in declarative sentences on the antepenultimate syllable of the sentence, in order to avoid the effect of undesired smooth decaying of Fo towards the final low tone. It can be considered a variant of the declarative terminal boundary tone. <ds> cannot be assigned either on a syllable that already carries an undeletable AG stress, nor on any of its adjacent syllables. Tone <ds> is assigned in all other declarative contexts, regardless of the word class and of the position of the syllable in the word. It deletes any deletable AG stress on either the same syllable (i.e. antepenultimate) or one of its neighbours. Thus <ds> is stronger than AG stresses in phonotactic adjustments. <ds> provides a significant improvement of the pitch pattern's perceptual quality. <rs> also avoids unconvenient Fo transitions : assigning <rs> is a way to maintain a flat low Fo pattern throughout long sequences of <-OPEN> words located between the IU boundary tone and the following AG stress. Yet some IU tones do not allow it.

The AG level adds one more binary distinctive opposition among tones : AG tones, as opposed to IU tones, have specific Fo pattern and location. Thus, including <ds> and <rs>, the complete set of available tones in the system provides 15 possibilities. Every undeleted AG stress is interpreted as a specific tone, where all variants of reduced, expanded range, rising or falling realisations may occur. Tone phonotactic rules determine the polarity of tone features responsible for this specification. At this stage, all remaining word initial AG stresses have been assigned a tone, but word final AG stresses get a tone (namely a downstep called <ab>) *only if* they belong to a question or to an exclamative sentence. Each tone assignment makes the vowel lose its <ASTRESS> feature. Thus after AG-parsing, only word-final AG stresses that have not been interpreted as a tone are still identifiable as AG stress, since the vowel still carries <ASTRESS>.

Retained <ASTRESS> feature is exploited in lengthening assignment. Vowels that belong to syllables that carry an IU boundary tone have feature <PSTRESS>, which is exploited in lengthening assignment too. In addition, the feature <STRESS> is assigned to all <+OPEN> word initial and final syllables (and also to <-OPEN> word initial and final syllables if they carry tone <ds>). Thus : vowels carrying an IU tone become <PSTRESS, STRESS>; <ASTRESS> becomes <ASTRESS, STRESS>; vowels carrying an AG tone (or a <ds> at word boundaries) become <-ASTRESS, STRESS>; other syllables are <-STRESS>. Lengthening rules are :
L1 if <-STRESS> (shortening rule)
L2 if <+STRESS,-ASTRESS>
L3 if <+STRESS,+ASTRESS>
L4 if <STRESS,PSTRESS>
L5 if sentence final <STRESS, PSTRESS>.

CONCLUSION:

This multilinear hierarchical model introduced significant improvements in the processing of linguistic knowledge. The autonomy of prosodic representations, of interpretative rules between levels, and of adjustments, avoids unsatisfactory prosodic patterns obtained directly from morpho-syntactic analysis.

REFERENCES :

[1] Carlson, R.; Granström, B. (1990), "An environment for multilingual text-to-speech development", Proc. of the ETRW on speech synthesis, Autrans, France, Vol.2, 73-82.
[2] Barber, S.; Granström, B.; Touati, P. (1988), "French prosody in a rule-based text-to-speech system", Proc. of 7th FASE Symp., 3, 967-974.
[3] Hirst, D. J. (1994), "The symbolic coding of fundamental frequency curves: from acoustics to phonology", International Symposium on Prosody, Yokohama, 1-5.
[4] Hirst, D. J. ; Di Cristo, A. (1984), "French intonation: a parametric approach", *Die Neueren Sprachen* , 83 (5), 554-569.
[5] Di Cristo, A.; Hirst, D. J. (forthcoming), "L'accentuation non-emphatique en français : stratégies et paramètres", *Hommage à I. Fonagy.*
[6] Pasdeloup, V. (1993), "A prosodic model for French TTS (...)", *Talking Machines : theories, models and designs*, Bailly et al. (Eds), Elsevier.