PROSODIC STRUCTURE REPRESENTATION FOR BOUNDARY DETECTION IN SPONTANEOUS FRENCH

Natalia Segal, Katarina Bartkova

France Télécom R&D, Lannion, France
Natalia.Segal@orange-ftgroup.com, Katarina.Bartkova@orange-ftgroup.com

ABSTRACT

Automatic speech processing has recently turned to the treatment of continuous spontaneous speech, which demands, among many other issues, a representation of its prosodic organization.

This paper presents a new approach to automatic prosodic boundary detection and prosodic unit structuring, based, with certain changes, on a descriptive theory of the French prosodic system initially proposed for prepared speech. This theory had been transformed into a set of rules so as to create a hierarchical representation of a phrase in spontaneous French in the form of a prosodic tree. The method had been manually verified and then applied to a spontaneous speech database in order to obtain a statistical description of prosodic structures.

Keywords: Prosody, intonation, spontaneous speech, speech segmentation, prosodic trees.

1. INTRODUCTION

The main purpose of this study was to find a representation for the intonation system of spontaneous French, and as its part, prosodic boundary detection.

Even if the intonation system of French, its phonetic and phonological specifics have been studied by many scholars and various theories were proposed [1, 3, 5, 6], there is still a great difficulty as to choose a good representation for spontaneous speech prosody.

This difficulty lays, firstly, in the lack of a unique description, and secondly and mainly, in the fact that proposed approaches to French intonation are mostly based and tested on prepared speech which is quite different in its nature from spontaneous speech used in our work.

At the same time, applied approaches to the spontaneous speech segmentation based on machine learning techniques do not sufficiently take advantage of a priori phonetic knowledge (e.g. [7]) and were not yet profoundly studied, especially for French. Few attempts generally concern quite limited applications [4].

All these considerations have led us to try to develop a representation for French spontaneous speech intonation based on a theoretical description for prepared speech, but with some necessary adjustments due to automatic data treatment and the specificity of spontaneous speech.

2. THEORETICAL BACKGROUND

In French, the main stress is principally marked by lengthening of the stressed syllable (which is at least two times longer than an unstressed syllable). It is always the last syllable of a rhythmic group (also called a prosodic group) which is lengthened, and it is usually also marked by a pitch movement (rising or falling). The division into prosodic groups is not completely voluntary, there being lexical and rhythmic constraints on the stress emplacement [5].

Melodic patterns help the listeners to structure the phrase and to decode the message. For example, a phrase can be interpreted as being interrogative or declarative according to its last pitch movement (rising or falling respectively). However, any phrase composed of more than one prosodic group will have a more complex prosodic structure, with a more or less prominent intonation pattern on every prosodic group.

According to the theory proposed in [5], which was used as our main framework, melodic movements on prosodic groups in French have no pre-established standard pattern but rather follow the two main rules:

- Amplitude of Melodic Variation rule (AMV)
- Inversion of Melodic Slope rule (IMS)

The mechanism of Amplitude of Melodic Variation marks two intonation units as being situated on the same or different levels. The higher a unit is in the prosodic structure of the phrase, the more significant its final pitch movement is supposed to be. The most prominent is the final

contour which indicates whether the announcement was intended as declarative or interrogative. We can thus attribute its level to every boundary: C0 – the highest level, phrase boundary, C1 – second most important boundary etc.

Inversion of Melodic Slope is the mechanism which causes the last stressed syllable on the boundary of lower level C_{k+1} to have a melodic slope direction opposite to that of higher level C_k. The lowest level pitch patterns are often phonetically neutralized (flat or slightly falling), as they have only to be differentiated from all other contours.

Thus, prosodic structure of any French phrase can be represented as a prosodic tree with particular values of prosodic parameters such as pitch slope and segmental duration on the final syllables of prosodic groups. Prosodic structure can vary to a certain degree according to speaker's intentions, but it is related to the syntactic structure of the phrase and is subject to constraints (stress clash and syntactic clash) [5].

We used these basic mechanisms developed for prepared French speech as a framework for the algorithm adapted to spontaneous speech.

3. METHOD OVERVIEW

3.1. Speech database used

The speech database contains the results of a customer satisfaction survey and is constituted of more than 1080 telephone messages in French. Since every message is pronounced by a different user, we are also able to approximate the number of speakers (male and female). Length of messages varies, with an average of 54 words per message. The database was manually transcribed in orthographic form with the annotation of non-speech noises as well as interrupted words and filled pauses. The orthographic representation was automatically given a phonemic transcription and aligned with the speech signal.

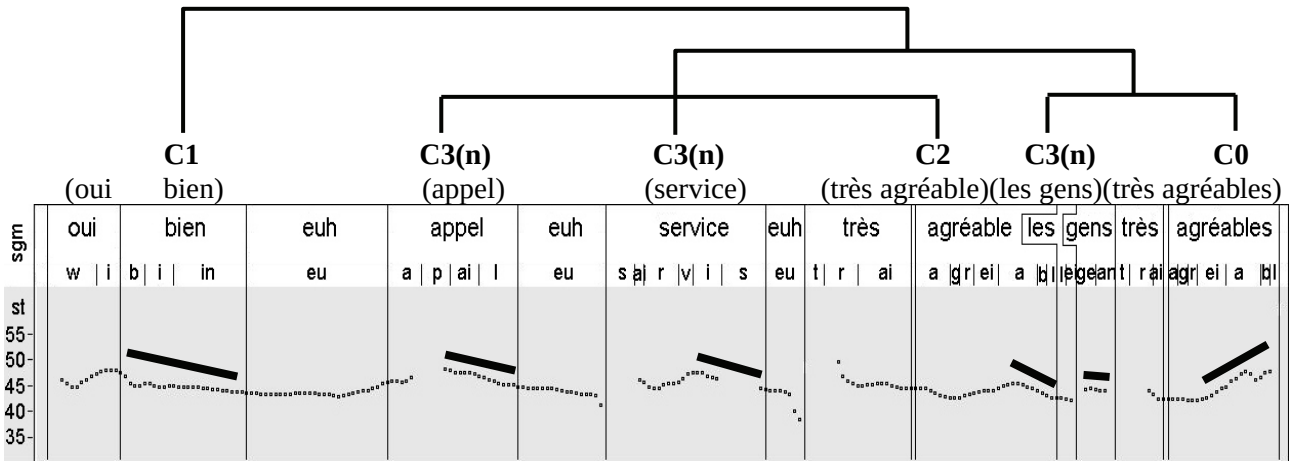
3.2. Prosodic trees

Using the aligned database described in previous paragraph, we built an algorithm to automatically produce prosodic trees for spontaneous speech. The construction of a prosodic tree is divided into three steps which closely follow the theoretical frame described previously.

3.2.1. Stressable words

As it was mentioned above, prosodic groups in French, though dependent on speaker's choice, are still subject to some constraints of lexical and rhythmic nature. To respond to lexical constraints, we designed a parser dividing a text into what we call stressable words, that is, groups of words that can potentially be stressed. The parser is based on 110 word categories for lexical and grammatical words, representing the class and function of every word. These word categories were yielded by the lexical module of our text-to-speech transcriber. Lexical markers are reunited into four "open" word classes: nouns, verbs, adjectives and adverbs. Words from open classes constitute the core of stressable words, grammatical words being attached as proclitics or enclitics.

Figure 1: Example of tree construction with irregular syntactic structure. Stressed words (parenthesized) are grouped together into trees according to the pitch value on their last syllable. The lowest level in the tree is neutralized (n): flat or slightly falling.



3.2.2. Stressed words (prosodic groups)

Our method determines, for each stressable word, whether it has been stressed, that is realized by the speaker as a prosodic group. Vowel duration has been considered here as the main parameter, since, as we have discussed above, last syllable lengthening is the main marker of stress in French. We have established a threshold for the ratio of the last vowel duration in a stressable word to the vowel duration of the following syllable. In particular case of a stressable word followed by a one-syllable stressable word which can also be accentuated, we compared the length of the last vowel of the stressable word not to the following vowel but to the mean vowel length of the phrase. We also have been taking into account pitch movement amplitude as a secondary parameter. Still, stressed words marked by an important pitch movement without duration lengthening were quite rare.

We also considered a rhythmic constraint by imposing a maximum of 8 syllables per prosodic group (stressed word). This maximum is reported to be either 7 or 8 [2], but 8 syllables produced better results on our data (most of the greater prosodic groups usually resulting from some treatment errors).

3.2.3. Tree construction

To proceed with tree construction phrase boundaries had to be detected. Defining what a phrase boundary is can be quite problematic in spontaneous speech because of many irregularities in word formation and syntactic structure due to hesitations, false starts, truncations and other disfluencies. That is why we decided to use only prosodic parameters values to determine the place of a phrase boundary, without any syntactical or lexical implications. In fact, phrase boundary is an important prosodic event and it may be expected to be sufficiently well marked by prosodic parameters. The main prosodic parameter used here is the amplitude of pitch variation (rising or falling) on the last syllable of a stressed word. A threshold has been determined for this amplitude on the basis of the manual segmentation of a part of the database. Long silent (unfilled) pause has been considered as another marker of phrase boundary. We didn't take into account filled pauses that usually mark hesitations and not phrase boundaries.

For each detected phrase we proceed with constructing its prosodic structure tree pursuant to the rules of IMS and AMV described in section 2. E.g., Figure 1 shows a phrase with the phrase boundary C0 marked by an important rising pitch movement. The second-level boundary C1 must then be marked by a slightly less prominent pitch movement of the opposite direction (falling) etc.

4. EVALUATION

The proposed method of segmentation and structuring of prosodic units was evaluated manually by an expert. This allowed us to measure the error rate at every step of the method, and to point out typical errors. We also gathered some statistical data on prosodic trees for the whole database and made conclusions on typical prosodic structures.

4.1. Manual tests

Manual evaluation was performed on 100 files of different lengths (from 1 to 56 phrases). The total number of phrases studied was 1333.

4.1.1. Stressed word detection

Table 1 represents the error rate for stressed word detection. Stressable word boundaries wrongly considered as stressed (inserted) and not recognized stressed word boundaries (omitted) were found by the expert.

Table 1: Efficiency of stressed word detection.

Detected	Inserted	Omitted	Recall (%)	Precision (%)	F_1-measure
2154	257	186	91	88	89.5

These errors were due to problems either with parsing into stressable words or with the algorithm for stressed words detection in itself. Errors of parsing were quite few (15) and did not affect the system performance much. As for the errors of stress detection, there were errors of alignment that changed vowel length and also many hesitations that can occur even on grammatical words and affect vowel length dramatically. Another reason is word emphasis that normally affects the first syllable of the emphasized word in French and changes its pitch but also its length.

4.1.2. Phrase boundary detection

Evaluating phrase boundary detection was much more troublesome since it is quite difficult, even for an expert, to tell the difference between "right"

and "wrong" detection for spontaneous speech. We tried thus to be as much tolerant as possible and accept all the boundaries that don't break minimal syntactic coherence inside a simple sentence and are not due to errors in the detection of parameters. Compound and complex sentences can be separated by phrase boundary, as any co-ordinate or subordinate connection. Table 2 summarizes the error rate for phrase boundary detection.

Table 2: Efficiency of phrase boundary detection.

Detected	Inserted	Omitted	Recall (%)	Precision (%)	F_1-measure
1327	139	145	89.1	89.5	89.3

It is interesting to mention that even an affirmative phrase in spontaneous speech can be marked by a rising pitch contour, which is considered to be a mark of an unfinished phrase in prepared speech. There were almost 1.5 times more rising contours than falling.

4.1.3. Tree structure verification

As it was discussed above, the prosodic tree structure is quite voluntary, and so in its evaluation we also tried to adhere to a tolerant position, verifying only that the tree doesn't break any syntactic constraint (there is no stress clash or syntactic clash). Of 1327 detected phrases, 186 had either a stress clash (72) or a syntactic clash (105) in its prosodic structure and were considered as errors.

4.2. Database statistics

After manual verification on a test part we applied the algorithm to the whole database.

4.2.1. General descriptive statistics

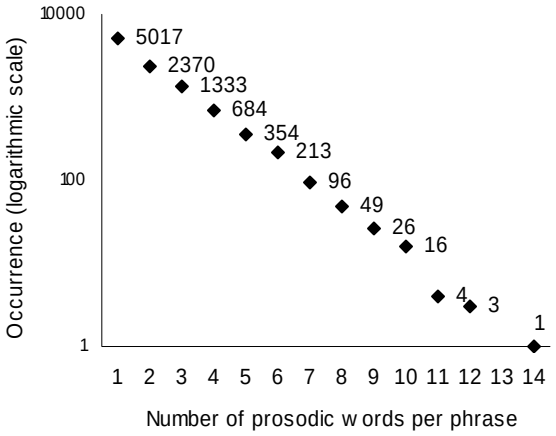
There were all in all 10166 phrases detected, their length varying from 1 up to 14 stressed words. The distribution of phrase length (in prosodic words) is given in Figure 2. As we can see, long phrases represent only marginal percentage of all the phrases and are usually due to some errors either in alignment or in parameters detection. In our analysis of prosodic tree structure we only considered the phrases no longer than 6 prosodic words that represent about 2.5 standard deviations (1.5) from the mean value (2.07).

4.2.2. Prosodic tree structures

Our general observation is that phrase structures tend to be rather balanced: trees usually prefer constructions where two nodes on one level include the same number of prosodic words.

Unbalanced structures favor the trees with most of its weight at the end of the phrase.

Figure 2: Length of prosodic phrase distribution.



5. CONCLUSION

We proposed an algorithm for automatic prosodic structure representation which proved to be adequate for spontaneous speech in French. The method allows to detect prosodic boundaries and to distinguish between them according to their level in prosodic tree. This permitted us to make a step forward in the understanding of the nature of prosodic boundaries. In perspective, we intend to use machine learning techniques to model different types of boundaries in prosodic trees.

6. REFERENCES

[1] Blanche-Benveniste, C. & al. 1991. *Le français parlé - Etudes grammaticales*. Paris: éd. CNRS.

[2] Fonagy, I. 1979. L'accent en français : accent probabilitaire. In: Fonagy, I., Léon, P. (eds), *L'accent en français contemporain*. Paris: Didier, 123-233.

[3] Hirst, D. 1998. Intonation in French. In: Hirst, D., Di Cristo, A. (eds), *Intonation Systems: a survey of twenty languages*. Cambridge: Cambridge Univ. Press, 195-218.

[4] Langlais, P. 1997. Estimating prosodic weights in a syntactic-rhythmical prediction system. *Proc. Eurospeech'97* Rhodes, 1467-1470.

[5] Martin, Ph. 1987. Prosodic and rhythmic structures in French. *Linguistics* 25, 925-949.

[6] Rossi, M. 1999. *L'intonation, le système du français : description et modélisation*. Paris : Ophrys.

[7] Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., Harper M. 2006. Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies. *IEEE Trans. Audio, Speech and Language Processing* 14(5), 1526-1540.