# Automatic Detection of Prosodic Cues for Segmenting Continuous Speech into Supralexical Units

Noëlle Carbonell, Yves Laprie
CRIN-CNRS & INRIA-Lorraine
BP 239, 54506 Vandœuvre-lès-Nancy Cedex, France

## ABSTRACT

*We are currently working on the automatic detection of syntagm and word boundaries in French, from the study of pitch and rhythm temporal evolution. After a brief presentation of our method, we discuss results obtained on various continuous speech corpora .*

## OBJECTIVES AND HYPOTHESES

Numerous experimental studies have brought out the useful contribution of prosody to speech perception and understanding (cf. for instance: Blesser, 1969 in Waibel, 1988; Grosjean, 1987). However, J. Vaissière and A. Waibel (Vaissière, 1988, Waibel, 1988) observe that few prototypes — among recent continuous speech recognition and understanding systems — take into account the linguistic information embedded in prosodic events.

The complex nature and multiple functions of prosody may account for this paradoxical situation, namely:

- the great intra- and inter-speaker variability that characterizes prosodic expression;
- the interactions between microprosody, which is deeply influenced by the phonetic contents of utterances, and macroprosody, which contributes to the interpretation of speech at various levels: lexical, syntactic and semantic (cf. Waibel, 1988, page 11);
- the complex relationships between macroprosody on the one hand, syntax and semantics on the other hand (Hirst et al., 1991).

We report a recent empirical study that we conducted on the contribution of French prosody to the automatic segmentation of continuous speech into lexical units. Our aim was to determine whether reliable predictions on word-boundary locations could be derived from the automatic analysis of pitch evolution and rhythm variations.

This issue has not yet been carefully investigated. In the few existing prototypes that involve prosodic data in the recognition/understanding process, prosody is assigned a limited function, that is: the validation of word hypotheses generated from acoustic-phonetic, syntactic or semantic evidence (Waibel, 1987).

But, in the case of French at least, the role of prosodic information should be carefully considered, since studies on French prosody (cf. in particular, Rossi, 1981; Di Cristo, 1981) indicate that: most often, main lexical accents are located on the last syllables of words, and last syllables of syntagms are associated with specific prosodic patterns.

Therefore, we contend — against (Vaissière, 1982) for instance — that the function of prosodic information in systems for the recogntion of spoken French should not, a priori, be restricted to the validation of word-boundary hypotheses generated by other information sources.

Our study is based on the following assumptions:

1. We hold, with M. Rossi and other phoneticians, that the contribution of energy to the expression of suprasegmental linguistic information is rather limited in French (Rossi, 1981), compared with the role that this prosodic parameter plays in other languages

(such as English, for instance); which explains why we have restricted the scope of our study to the analysis of pitch evolution and rhythm variations.

2. We assume, with I. Gaïtella (Gaïtella, 1991) that prosodic expression greatly varies according to the pragmatic context (or situation) in which speech is produced; therefore, our study includes the analysis of speech corpora recorded in various situations (reading, sentence repetition from memory, simulated human-computer dialogue).
3. We chose to use macroprosody as an independent source of information, in order to promote robust speech recognition and understanding; therefore, we decided to focus pitch analysis on the centre parts of vocalic nuclei exclusively, so as to eliminate most microprosodic effects; with regard to the study of rhythm, it is based on a rough estimate of the mean duration of vocalic nuclei per prosodic group.
4. Since phoneticians do not agree on a unique model for the description of French macroprosodic sentence patterns, we did not attempt to interpret prosodic cues in terms of comprehensive macroprosodic structures; instead, we focussed on the detection of prosodic cues that may contribute efficiently to lexical segmentation, that is: syntagm-boundary markers and chief lexical accents; with respect to the latter, we assumed that they could be reliably distinguished from secondary and emphatic accents which may occur on first or intermediate syllables in words (cf. polysyllables).

We briefly describe our method in the next paragraph. Then, results obtained on three different speech corpora are presented and discussed. Finally, we indicate how these results influence our current work on the contribution of prosody to speech recognition.

## METHOD AND ALGORITHMS

### Pitch evolution analysis

In order to get robust pitch data, we chose to implement the time-domain auto-correlation method for pitch detection proposed by M.M. Sondhi in 1968 and improved by L.R. Rabiner namely.

We developed an algorithm for detecting peaks on the F0 curve (in the time x frequency plan), since in French:

- F0 values are significantly higher on stressed syllables (lexical accent) than on unstressed syllables,
- last syllables in some syntagms correspond with prominent local maxima on the F0 curve (cf. the notion of 'continuation majeure' in Rossi, 1981).

Peaks are determined on the smoothest curve (natural cubic spline) resulting from interpolating selected F0 values: one per vocalic nucleus. With respect to syntagm-boundary and stress detection, results are best when the selected values are located in the centre parts of vocalic nuclei; which is in keeping with the observation that microprosody effects are reduced on vowels.

Vocalic nuclei are identified thanks to a speaker-independent algorithm (NOVOCA) (Fohr, 1989); decision criteria are based on a coarse analysis of the spectral distribution of energy.

### Study of rhythm variations

Our goal was to detect rhythm variations that contribute to the marking of prosodic group endings and the expression of stress. In French, stressed syllables and last syllables in most syntagms (Di Cristo, 1981) are significantly lengthened.

Since present acoustic-phonetic decoders cannot reliably segment speech into syllables, we used vocalic nucleus duration as a basic unit for the study of rhythm, although vowel duration is greatly influenced by the nature of the vowel and by its phonetic context (1). Our choice is a compromise between practical considerations (i.e. the limits of present acoustic-phonetic decoders) and the constraint (cf. assumption 3 in the previous

(1) For instance, a nasal vowel is often significantly longer than an oral one. As an illustration of context effects, /a/ in syllables where it is followed by /R/ (cf. "part" for example) is longer than when it is followed by another consonant as in "chatte".

paragraph) that prosody should be used as an independent information source; thanks to the accuracy of NOVOCA (with respect to vocalic nucleus detection and bounding), this constraint is not infringed.

In order to determine significant vocalic nucleus lengthenings, we compare each vocalic nucleus in a speech fragment with the Mean Vocalic Duration (MVD) calculated over this speech fragment (2). If the length of a vocalic nucleus is superior to 1.5 x the MVD, we consider that it has undergone a significant lengthening.

## PRESENTATION AND DISCUSSION OF RESULTS

### Speech corpora

We have tested our algorithms on three multi-speakers speech corpora recorded in three different speech production pragmatic environments:

- corpus LABISE: reading of a short text in a sound-proof room (12 male speakers);
- corpus CMB: 5 male speakers in a quiet room were asked to say rapidly short sentences from memory (short-term memory);
- corpus METEO: simulated human-computer oral dialogues (on weather-forecast) in a realistic environment: each speaker (10 male speakers on the whole) interacted with a real microcomputer to which he could also 'talk' spontaneously.

### Results

Results are summarized in the following table.

**Table 1.** *Automatic detection of macroprosodic cues (i.e. F0 peaks and vocalic nucleus lengthenings) for lexical segmentation — Three speech production situations are considered: reading, 'aloud recollection' of sentences, quasi-spontaneous dialogue.*

*nb. syntagms: number of potential syntagmatic marks ('continuations majeures').*
*correct (F0): detected F0 peaks corresponding to syntagm endings.*
*lexical (F0): detected F0 peaks corresponding to word endings (inside syntagms).*
*errors (F0): detected F0 peaks located on first or intermediate syllables in words.*
*correct (D): detected vocalic nucleus lenghtenings associated with syntagm endings.*
*errors (D): detected voc. nucl. lengthenings on first or intermediate syllables in words.*

*Percentages are calculated in reference to:*

- *the number of potential marks (first column), for the second and fifth columns;*
- *the number of correct lexical detections (syntagm+word marks), for columns 3, 4, 6.*

| Corpus | nb. syntagms | correct (F0) | lexical (F0) | errors (F0) | correct (D) | errors (D) |
|---|---|---|---|---|---|---|
| LABISE | (479) | **78%** | 19% | *9%* | **19%** | *6%* |
| CMB | (159) | **88%** | 7% | *8%* | **30%** | *12%* |
| METEO | (1282) | **83%** | 26% | *10%* | **25%** | *18%* |

### Discussion

Results in table 1 indicate that our algorithms are capable of detecting 4 out of 5 syntagm boundaries, thanks to the analysis of F0 variations only. Errors (i.e. erroneous word-boundary detection) are due to:

- optional secondary lexical accents in polysyllabic words with 3 or more syllables,
- and expressive/emphatic effects, which take the form of lexical accents but are located on first or intermediate syllables in words (Aouizerat and Lonchamp, 1991).

Since error rates are inferior or equal to 10% in the three conditions, word-boundary marks obtained from the analysis of F0 macrovariations may contribute efficiently to the validation of lexical hypotheses generated from other information sources in speech recognition systems. But better detection rates are needed, in order to ensure the success of

(2) We designate by "speech fragment" any sequence of speech sounds bounded by two successive pauses (i.e. silences of 250 ms or more). For any given speech fragment, the MVD is the median value in the set constituted by the durations of all vocalic nuclei included in this speech fragment.

attempts aiming at improving lexical identification thanks to the generation of word-boundary hypotheses from melodic cues.

Besides, syntagm-boundary detection is unreliable, especially regarding spontaneous speech and reading (cf. column 3 in table 1). Then, results from our algorithm should not be used at the syntactic level: their contribution might drastically reduce the accuracy and efficiency of syntactic analysis.

Results from the study of rhythm are rather disappointing. Error rates are high, especially for spontaneous speech. Moreover, relatively few syntagm boundaries are correctly detected: 1 out of 5 for reading, and 1 out of 4 for spontaneous speech. Besides, accuracy greatly varies from one pragmatic situation to another; differences between reading and spontaneous speech are particularly marked. Which suggests that, in French, intonation conveys more linguistic information than rhythm, and that rhythm is more influenced (than intonation) by other factors, such as pragmatic constraints on speech production.

Therefore, information supplied by the automatic analysis of rhythm cannot be involved in the interpretation of French spoken utterances, at least for the time being. Further studies are necessary, in order to determine the exact contribution of rhythm to speech interpretation.

## CONCLUSION

The empirical study on French prosody that we have presented here points to the following conclusions. First, pitch analysis appears as a useful source of information for validating lexical segmentation hypotheses generated by word recognizers that operate on continuous speech, which confirms conclusions from previous studies. Secondly, the contribution of rhythm to lexical segmentation (more generally, to speech interpretation) is not clear, and requires further investigation.

Moreover, our results suggest two research directions that we are currently investigating:

- the definition and testing of criteria for assessing the reliability of prosodic analysis results, with a view to selecting the most robust cues for the generation of word-boundary predictions;
- the refinement of pitch and rhythm analysis, in order to improve syntagm-boundary detection; which implies the interpretation of F0 evolution and syllabic duration variations in terms of macroprosodic groups and patterns.

## REFERENCES

Aouizerat H., Lonchamp F. (1992), "Description et génération par règles de l'intonation de la phrase énonciative lue en français", *Séminaire Prosodie*, Aix en Provence, Octobre 1992.

Di Cristo A. (1981), De la microprosodie à l'intonosyntaxe, Thèse d'Etat, Aix en Provence.

Fohr D., Carbonell N., Haton J.-P. (1989), "Phonetic Decoding of Continuous Speech with the APHODEX Expert System", in J.-P. Tubach, J.-J. Mariani (eds.), *Proc. EUROSPEECH 89*, Paris, Septembre 1989, Vol. 2, pp. 609-612.

Gaïtella A. (1991), Rythme et parole: comparaison critique du rythme de la lecture oralisée et de la parole spontanée, Thèse d'Université, Aix en Provence.

Grosjean F., Gee F. (1987), "Prosodic structure and spoken word recognition", *Cognition 25*, pp. 135-155.

Hirst D., Espesser R., Di Cristo A. (1991), "Constituants prosodiques et macro-segmentation du signal: méthodologie et critères d'évaluation", *Actes Séminaire Prosodie et Reconnaissance de la Parole*, Aix en Provence, Mars 1991, pp. 1-8.

Rossi M. (1981), *L'intonation - De l'acoustique à la sémantique*, Paris: Klincksieck.

Vaissière J. (1982), "A Suprasegmental Component in a French Speech Recognition System", *Recherches Acoustiques*, CNET, Vol. VII, pp. 109-125.

Vaissière J. (1988), "The use of Prosodic Parameters in ASR", in *Recent Advances in Speech Understanding and Dialog Systems*, Berlin: Springer-Verlag, pp. 71-99.

Waibel A. (1987), "Prosodic Knowledge Sources for Word Hypothesization in a Continuous Speech Recognition System", *Proc. IEEE ICASSP-87*, Dallas, Avril 1987, pp. 856-859.

Waibel A. (1988), *Prosody and speech recognition*, Londres: Pitman.