On the Use of Prosodic Knowledge for Continuous Speech Recognition and Understanding

N. CARBONELL
CRIN-INRIA Lorraine
B.P. 239, F-54506 Vandoeuvre lès Nancy Cedex France.

Abstract.
We are currently studying how prosodic cues could facilitate continuous speech recognition and understanding. In this paper, results are discussed concerning the detection of syntagm boundaries from the analysis of fundamental frequency and rhythm variations. Three corpus have been tested corresponding to different speech production conditions: reading, memorized sentences (short-term memory), quasi-spontaneous information dialogues. Our present rates for the detection of word or syntagm boundaries are roughly: 90% correct markers.

Although numerous authors assert the usefulness of prosodic information for the recognition of large vocabularies or the understanding of continuous speech, few attempts have been made yet, in order to implement prosodic knowledge and integrate it into speech recognition/understanding computer systems.

This paper reports an attempt to involve prosodic information in the recognition/understanding of continuous speech, i.e. French sentences spoken naturally. We first present our goals and the exact scope of our study. Then, we briefly survey the numerous difficulties raised by the detection of prosodic knowledge useful for improving continuous speech recognition. In a third section, we describe our study and discuss implementation results; our approach is presented together with the speech corpus and signal processing techniques used.

Finally, we indicate how prosodic information can ameliorate continuous speech recognition; more precisely, we attempt to define the exact nature and limits of the role that prosodic information can play in continuous speech recognition and understanding systems; we also try to determine how such information can cooperate with other information sources so as to improve recognition accuracy.

1. OUR OBJECTIVES.

Phoneticians agree that time variations of prosodic parameters, especially the fundamental frequency (Fo) and syllabic duration, have linguistic significance. Consequently, it may be assumed that, by taking into account prosodic information, one may greatly improve the results of continuous speech recognition and understanding systems (CSRUS). Which raises the following crucial question: what role should prosodic information play among other information sources in a CSRUS?

The view advocated by W. A. Lea in the seventies [12], namely that prosodic information can and should guide the recognition and understanding process, has never been carefully investigated and was abandoned long ago, since it rapidly proved unrealistic. Most recent systems based on the cooperation of several knowledge sources include a prosodic component besides acoustic-phonetic, lexical, syntactic, semantic and pragmatic modules [11, 13, 15, 16, 21, 23]. Nevertheless, prosodic information is often assigned a limited and subsidiary function in the process of continuous speech recognition/understanding (CSRU), mainly because prosodic cues are very ambiguous and difficult to detect reliably (due to intra- and inter-speaker variability). Moreover, a few algorithms only, such as those developed by J. Vaisstère and P. Martin, take into account all prosodic parameters, i.e.: Fo, rhythm and global energy. Most often, Fo only is considered.

The chief aim underlying the study reported here is to derive meaningful prosodic cues from the analysis of Fo macro-variations, vocalic duration fluctuations and pauses, in order to improve CSRU and to integrate prosodic information and knowledge into a system (DIAL) for the understanding and managing of voice task-oriented dialogues using natural language, which is being developed by our research group at CRIN [4]. Global energy has not been considered, since this prosodic parameter is much less important in French than in other languages and has been less studied by phoneticians than Fo or the rhythm.

Before describing our approach, we shall bring out the difficulties inherent to the detection of prosodic cues in continuous speech and to the use of prosodic information in CSRU.

2. DETECTION AND INTERPRETATION OF PROSODIC CUES IN CONTINUOUS SPEECH.

Most difficulties arise from the fact that the variations of prosodic parameters (Fo, rhythm and energy) are highly ambiguous, since they result from multiple heterogeneous factors [22], such as:

- the nature of the sequence of phonemes uttered and coarticulation acoustic effects, which induce micro-variations of prosodic parameters (microprosody);
- the syntactic structure of the spoken statement, its semantic content and the speaker's attitude (cf. emphasis phenomena), which are responsible for macro-variations (macroprosody).

Intra- and inter-speaker variability also complicates the detection and interpretation of prosodic cues.

We illustrate these difficulties with a few examples taken from a multi-speaker corpus of quasi-spontaneous continuous speech, which is described in paragraph 3.1 (corpus METEO):

- an emphatic stress [speaker's attitude] may induce, on the first syllable of a polysyllabic word, a Fo "peak" (i.e. a successive rise and fall) similar to the Fo "peak" that, in French, is usually associated with the last vowel in syntagms which do not end a sentence [syntactic structure]; for instance: "...les températures maximales et minimales..." vs. "...les températures maximales aujourd'hui..."
- in French, vowels followed by certain consonants are lengthened [microprosody] as well as stressed vowels and vowels at the end of syntagms [syntactic structure] so that, for a given speaker, the relative duration of a phoneme occurring in an unstressed position within a syntagm may be equal to the duration of an occurrence of the same phoneme at a syntagm boundary, i.e. 1.5 x the mean vocalic duration at least; for example, we have observed in the speech corpus from one of our speakers that, in "... pager Remiremont..." and "... les températures minima, aujourd'hui...", both [a] have the same duration (2 x the mean vocalic duration); the first [a] is lengthened because it is followed by [R] in the same syllable [intra-speaker variability and microprosody], whilst the lengthening of the second one is due to its position at a syntagm boundary [syntactic structure];
- some speakers emphasize meaningful rhythm variations; others, on the contrary, do not significantly increase the duration of stressed syllables or syllables ending syntagms [inter-speaker variability]

3. OUR STUDY.

3.1. Speech corpus used.
In order to test and validate the various algorithms we have developed for the detection and interpretation of prosodic cues, we have collected three speech corpora corresponding to three different speech production conditions: reading, quasi-spontaneous speech, and an intermediate condition akin to dictation.
We have adopted this strategy on the basis of the following assumptions:
- the nature and form of meaningful macroprosodic events vary with speech production conditions; in other words, the prosody involved in reading or dictation greatly differs from the prosodic characteristics of continuous speech.
We now briefly describe the speech data on which the quantitative results given in paragraph 3.3 are based.

**Common characteristics of the speech corpus:**
- Sampling frequency: 16 kHz.
- Untrained, unselected speakers.
- Phonetic segmentation and labeling by expert phoneticians, from high quality numerical speech spectrograms [11]; for the sake of coding homogeneity, each corpus has been processed by only one phonetician.

**Read speech corpus (LABISE):**
A relatively simple text called "La bise et le soleil" was read by fifteen male and fifteen female speakers; high quality recording has been achieved thanks to the support of the CNRS-GRECO "Communication parlée".
At present, this corpus is but partly segmented and labeled; at present, the speech from eleven speakers has been processed by the expert. Therefore results bear on the labeled part only.

**Quasi-spontaneous speech corpus (METEO):**
Simulated voice human-computer information dialogues were recorded in a noisy environment (computer room).
Ten male speakers were asked to request specific information relating to weather forecast by dialoguing with an Automatic Meteorological Information Center simulated by a human operator [3]. No speech constraint was imposed on speakers.

**Intermediate condition (corpus CMB):**
Fifty different sentences were uttered by five male speakers (ten sentences per speaker) according to the following protocol: the text of each sentence selected from the corpus of phonetically balanced sets of sentences (ten sentences per set) designed by P. Combescure [5] is given to the speaker who reads it silently, memorizes it (short-term memory) and then speaks it as naturally as possible.
Recording took place in a relatively quiet room.

3.2. Speech analysis software developed.

**Fo estimation:**
We have chosen the autocorrelation method initially proposed for the determination of Fo by M. M. Sondhi [20] because of its robustness in noisy environments especially. More precisely, we have implemented the algorithm described in [18] which performs center and infinite clipping on the speech signal before evaluating the autocorrelation function.
Regarding the detection and correction of erroneous Fo values, we have refined the algorithm designed by G. J. Bristow and F. Fallside [2], in particular by taking into account the average Fo value calculated over the whole speaker's utterance.
An efficient implementation (a few seconds processing time) of these algorithms has been developed in C by J.-J. Bonin [1] on a Masscomp S600 (Real Time Unix / array processor). Results are shown on Figure 1.

**Vocalic nucleus detection and "mean vocalic duration" evaluation:**
The detection of vocalic nuclei is based on the analysis of the temporal distribution of energy within an appropriate frequency band. We call "mean vocalic duration" the median value among all vocalic nucleus durations in the current utterance; both algorithms are described in [9].
The notion of mean vocalic duration helps us to determine meaningful rhythm variations, whilst the detection of vocalic nuclei enables us to restrict the scope of our analysis of Fo variations to relatively stable parts of the signal (vocalic nuclei centers), in order to eliminate microprosodic effects.

**Detection of meaningful Fo peaks:**
Fo peaks are located on the curve constituted by Fo values chosen as follows: for each vocalic nucleus in the utterance, we select the maximum Fo value in the nucleus. The algorithm developed by J.-J. Bonin [1] first smooths the curve, then determines major peaks thanks to flexible contextual thresholds and finally positions syntagm boundary markers at the vertices of the selected peaks. An example is given in Figure 1. This algorithm too has been programmed in C and is running on our Masscomp.

---

![Figure 1: Fo curve and syntagm boundary markers for the sentence](image)

"J'aime(e)rais connais(lle) le temps prévu le douze Juin mille neuf cent quatre vingt deux." (METEO)
We are now thinking of refining our approach by taking into account Fo variations inside vocalic nuclei ; we hope that such information will prove useful, at least for discriminating stressed syllables from syllables ending syntagms, although it may be combined with and masked by the effects of microprosodic phenomena.

3.3. Results.

Present results concern the detection of syntagm boundaries from the interpretation of Fo and vocalic duration variations. Information especially has been carefully considered, since it has been much studied by phoneticians and various melodic patterns and models have been proposed so as to account for observed successions of Fo rises and falls within a sentence [19, pp. 179-235], [14], [10].

We have chosen to deduce fragmentary but reliable robust structural information from prosodic cues rather than to elaborate complete complex melodic representations of sentences in terms of the available models, for the main following reason:

the construction of a global melodic interpretation for each sentence involves a careful analysis of Fo variations inside syllables ; meaningful acoustic cues that can be easily and reliably detected on the speech signal are yet to be found, since the evolution of Fo inside syllables is often ambiguous (cf. section 2) ; the interaction between micro- and macro-melodic phenomena especially, may cause numerous erroneous melodic interpretations ; acoustic-phonetic and syntactic-semantic information is necessary to solve ambiguities.

At present, our algorithms are capable of positioning some syntagm boundary markers, from the detection of prosodic patterns associated with syntagmatic boundaries ; the notions of "continuation majeure" and "continuation minueure" which have been characterized by P. Delattre [6] have been taken into account, whilst "glissandos" which have been defined by M. Rossi [19] have been ignored because they are difficult to detect ; the height of vocalic nuclei that has been studied by P. Delattre [7], M. Rossi [19] and A. Di Cristo [8], has been used as an additional confirmatory cue.

Our next step will be to produce complete prosodic descriptions of sentences, once we have defined appropriate intra-syllabic acoustic cues ; such information, although difficult to obtain, may prove very useful for the recognition of continuous speech, especially natural language statements.

Description of results : Results relating to the three corpus have been grouped in the table of Figure 2 ; columns contain counts of the following items for the three corpus : - column 1 : potential syntagm boundary markers, i.e. syntagm boundaries that speakers may bring out thanks to prosody ; we have applied M. Rossi's intonation model for the determination of potential prosodic markers [19, p. 203] ; - column 2 : correct syntagm boundary markers found by our algorithms ; - column 3 : errors made by our algorithms, for instance : markers positioned on the first syllable of a polysyllabic word, due to lexical or emphatic stress ; - column 4 : lexical boundaries detected by our algorithms correctly.

Percentages are defined : - for the second column, in relation to the first one, - and for the third column, in relation to the sum of the second and the fourth columns.

<table>
<thead>
<tr>
<th>corpus</th>
<th>potential correct syntagm markers</th>
<th>errors</th>
<th>correct word marker</th>
</tr>
</thead>
<tbody>
<tr>
<td>LABISE</td>
<td>450</td>
<td>159</td>
<td>141 88%</td>
</tr>
<tr>
<td>CMB</td>
<td>1282</td>
<td>1065 83%</td>
<td>151 10%</td>
</tr>
<tr>
<td>METEO</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: Table of results.

Comments : More than four out of five of the potential syntagm boundary markers in our corpus are correctly detected by our algorithms. Differences between corpus may be explained as follows : - macroprosodic phenomena are less frequent and less conspicuous in read texts than in more spontaneous speech ; this observation may account for the rather poor results obtained with the LABISE corpus ; - results relating to the CMB corpus are significantly superior to those concerning the METEO corpus perhaps because, in spontaneous speech, syntactic constraints are relaxed, and prosody, besides its expressive function, is used to enhance the semantic content of utterances rather than their syntactic structure.

Error rates are low, about 10% of all the word boundary markers provided by our algorithms ; they do not significantly vary from one corpus to another.

Most errors arise from confusions between prosodic manifestations of lexical or emphatic stress and prosodic phenomena associated with syntagm boundaries ; for instance : - lexical stress : "On entend le gazouillis d'un oiseau dans le jardin." (CMB) - "Quelle sera la température maxima à Villers..." (METEO) - emphatic stress : "J'aimerais faire le temps prévu à Nancy demain." (METEO)

4. HOW TO USE PROSODIC INFORMATION IN CONTINUOUS SPEECH RECOGNITION.

We are considering the integration of our prosodic analysis module into the voice dialogue managing system DIAL mentioned in section 1.

Figure 3 (cf. [17]) describes the architecture of DIAL and precis the type of information exchanged between the prosodic component (PROSO) and the other processors in the system.

More precisely, PROSO detects syntagm boundaries from the speech signal and from a few data produced by the acoustic-phonetic decoder (APHON), that is : the position and boundaries of the vocalic nuclei in the utterance on the one hand, and, on the other hand, the mean vocalic duration. This module provides the lexical component (LEX) with word boundaries and the syntactic-semantic analyzers (SYN-SEM) with syntagm boundaries.

We assign to PROSO the following two following functions : - to validate/confirm hypotheses submitted by any other component ; in most knowledge-based speech understanding prototypes using multiple information sources, the only function of the prosodic component consists in verifying hypotheses elaborated by other modules ; - but also, more ambitiously and originally, to provide LEX and SYN-SEM with hypotheses on word/ syntagm boundaries.

This approach is consistent with the flexible recognition/understanding strategy that has been implemented in the DIAL system which can switch from top-down to bottom-up local analysis, according to the nature of current available information [4]. But our present results concerning word and syntagm boundary detection (10% error rate) should be significantly improved, in order that hypothetized prosodic boundaries can be used efficiently as anchor points for lexical matching or syntactic analysis. This present lack of accuracy explains why most existing CSRSUs restrict the role of prosody to the verification of hypotheses submitted by other modules in the system.

Besides, in order to improve recognition accuracy, PROSO should not deduce its interpretations from hypotheses produced by other components in the system. More precisely, since lexical and syntactic-semantic hypotheses are based on acoustic-phonetic hypotheses mainly, PROSO and APHON must process different acoustic cues, and PROSO must avoid using information elaborated by LEX and SYN-SEM, in order that prosodic information may help to correct interpretations produced by other components, and increase the quality and robustness of recognition ; otherwise, PROSO cannot be used for correcting or confirming hypotheses from LEX and SYN-SEM.

This explains
5. CONCLUSION.

Our study shows that the knowledge acquired by phoneticians on intonation and rhythm may contribute to improve the quality of speech recognition.

In particular, it is possible to design a robust reliable algorithm for the detection of syntagm boundaries; such an algorithm may also detect some lexical boundaries.

Present error rates, although low, should be reduced if we want to use prosodic information for hypothesizing word and syntagm boundaries, as well as for testing hypotheses produced by the syntactic-semantic and lexical modules in a CSRSUS.

The results of the perceptive analysis that we have performed on the METEO corpus [3] clearly indicate that listeners can perfectly well distinguish stress from prosodic effects at the end of syntagms. It then seems possible to improve the accuracy of our syntagm boundary detection algorithm, since most of its errors result from confusions between syntagm boundaries and lexical or emphatic stress on polysyllabic words.

But, in order to disambiguate stress from syntagm boundary prosodic marking, it is necessary to analyze intra-syllabic F0 variations; and it is difficult to isolate prosodic intra-syllabic effects related to the syntactic structure of the current utterance from microprosodic effects due to coarticulation, without knowing the nature of the phonemes that have been uttered, hence without using the phonetic lattice generated by the acoustic-phonetic decoder in a speech recognition system.

Therefore, we are now trying to find new acoustic cues that could help to disambiguate stress from syntagm boundaries. Besides, we are considering using the phonetic lattice produced by APHON in order to isolate meaningful F0 variations inside syllables and, consequently, to characterize F0 curves at syntagm boundaries in terms of the syntactic-prosodic typology proposed by phoneticians.

If we could significantly improve the quality of syntagm boundary detection and characterization, it would be possible to extend the role of prosodic information and knowledge in a CSRSUS; the functionalities of the prosodic component would no longer be restricted, as in [21], to the validation of hypotheses submitted by other components in the system or, as in DIAL [4], to the production of some word/syntagm boundary hypotheses and the validation of these hypotheses from other components, using specific information rather than conclusions produced by the other components in the system.

It could function in the same manner as the other components, that is: besides the validation of punctual hypotheses from other modules, it could generate complete specific representations of utterances thanks to interactions and exchanges of information with the other processors in the CSRSUS.

REFERENCES

[16] Perennou G., Caelen G., "Utilisation de la prosodie pour la reconnaissance de la parole dictée", in [24], pp. 25-57.