PROSODY AND SPEECH RECOGNITION : EXPERIMENTS AND PERSPECTIVES

Philippe Langlais

Laboratoire d'Informatique d'Avignon
Centre d'Enseignement de Recherche en Informatique
339, chemin des Meinajariès
BP 1228 — 84 911 F-AVIGNON Cedex 9
Tél. : (+33) 04 90 84 35 25 — Fax : (+33) 04 90 84 35 01
E-mail : langlais@univ-avignon.fr

Abstract

This paper concerns the study of information derived from the melodic, temporal and intensity characteristics of the material to be recognized in a speech recognition system, in French. One classical method for automatic prosodic analysis consists of three steps : parametrization, normalisation of the raw data taking into account the identity of the segments, and perception, and the application of suprasegmental rules.
Part I concerns the intrinsic and co-intrinsic variations. Are all these microfluctuations detectable by automatic techniques, and if they are, can they be used to improve an acoustic-phonetic decoding process ? Our study shows that only few phenomena can be significantly observed by means of these techniques on large isolated French words corpora. The reliable phenomena have been successfully integrated into a lexical access system.
Part II concerns the applications of suprasegmental rules to speech recognition. An automatic correlative system enabling experts to perform prosodic analysis has been developed (by providing them with query and vizualisation tools). It also enables to automatically put forward valued hypotheses concerning the linguistic organization of a message to decode.
Two applications of this system are proposed : the first one for the recognition of decimal numbers (our system is able to locate the word "virgule" (point) in an unknown number, only by means of prosodic information) and the second one for the recognition of French sentences read separately. The results obtained fully validate our approach.

1 Introduction

Despite a large number of studies concerning French prosody, and a fairly good understanding of the domain, it has not yet been possible to exploit such knowledge in an automatic speech recognition (ASR) system [Di Cristo, 75, Di Cristo, 81, Rossi, 81, Rossi, 85, Mertens, 90, Caelen, 91, Rossi, 93]. This relative failure is mainly due to three kinds of difficulties :

- an important contextual variability of prosodic knowledge (type of speech, speaker, structure and content of sentences, nature of the environment, etc.) ;
- the complexity of relations between prosodic information and various linguistic organization levels of a message ;
- the problems encountered with the accurate measurement of prosodic parameters, and their possible integration on a perceptual level.

Attempts to integrate prosody into ASR were common in the early eighties [Méloni, 82, Carbonell, 82, Vaissière, 82, Vaissière, 84], when French researchers were trying to integrate phonetic knowledge into general ASR systems. This work was marginal compared to research in phonetic decoding and speech modeling. Prosodic research for ASR has become even more marginal as statistical models have become more successful. Recent progress in ASR encourages the use of spontaneous speech, and there is a renewed interest in the use of prosody in ASR, as prosody plays a very important role in spontaneous speech in which grammatical constraints are loosened.
Past studies presented possibilities for integrating prosody into ASR systems [Lea, 80, Vaissière, 88]. This paper discusses the reliability of such integration in light of the experiments carried out on corpora of significant sizes.

2 Microprosody

Di Cristo'work [Di Cristo, 78] has undoubtedly contributed the most to highlighting the essential role of microprosodic variations in the French language. These fluctuations have also been highlighted by other authors in other languages [Lehiste, 76]. According to Lehiste, microprosody is defined as the intrinsic and co-intrinsic influences of fundamental frequency (F0), duration and intensity, due to segment identity and to phonetic context. The experiments presented in this study have been motivated by two reasons :

- some segmental configurations of prosodic parameters can prove to be relevant indexes in an acoustic-phonetic decoding module [Vaissière, 88, Dumouchel, 94] ;
- microprosodic variations are not motivated from a linguistic point of view and therefore disturb suprasegmental analysis.

To ensure the reliability of microprosodic information in an ASR context, it is necessary to validate this approach on large corpora with our automatic parameter extraction techniques. We know that prosodic parameters are governed by multiple factors. For instance, according to Klatt [Klatt, 75], there are at least 7 factors which influence segment duration : the word-, phrase- and sentence stresses, the emphasis of the word to which the phoneme belongs, the previous and following phonemes, the position in the sentence, the position in the word and finally, the nature of the segment.

The difficulty of a multi-parametric corpus analysis, as well as the quantity of necessary data to validate it can easily be imagined. A first study of the microprosody of isolated words has been carried out, thus eliminating supralexical factors governing the parametric variations. We shall therefore consider only the information coming from the intrinsic and co-intrinsic variations of prosodic parameters.
Several multi-speaker corpora of isolated words have been used in this study (for a total of 5600 words uttered by 7 different speakers). One of them (called FeLex) is made of 800 trisyllabic words ($\times$ 2 speakers) selected so as to achieve homogeneity in the number of representatives of high, low, mid and nasal vowels, in the right and left, voiced and unvoiced consonantic context for the three syllabic positions (beginning, middle and end of word). These words have been selected from Bdlex lexicon [Pérennou, 92] so as to obtain a number of vowels that approaches 100 for each context.
Our experiments aim to check whether the distributions of duration, pitch and intensity measures from the corpora allow us to make decisions (in the Bayesian sense) about the nature of the different segments observed.

2.1 Duration

The precise segmentation of a sentence into phonemes is a delicate task, even if performed manually. Campbell [Campbell, 92] points out that this is certainly a challenge and recommends the use of Markovian phonetic models. Two techniques of phonetic unit duration measurement have been investigated in this study :

- the lexical access module developed at the Avignon laboratory (called SPEX) [Béchet, 94] provides a temporal alignment for each recognized word. A block diagram of this system is provided in figure 1.
- non contextual phonetic models trained on different telephone multi-speaker databases recorded at IDIAP[1] [Chollet, 95]. The basic acoustic analysis used an observation vector consisting of 12 Mel frequency cepstral coefficients and log energy, 13 first-order differential coefficients and 13 second-order differential coefficients. A set of 38 models, with 3 states each (left-to-right topology, multiple mixture, full covariance matrix) has been defined with the HTK toolkit [Young, 90]. The examination of the alignment provided by these simple models in "problematic" contexts (ex : voiced consonant) has allowed to check the reliability of their behavior.

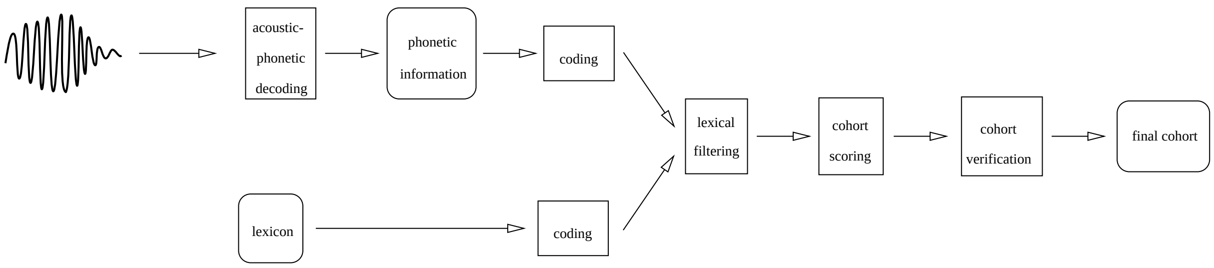

Figure 1: *Block diagram of the lexical access architecture of SPEX.*

[1]Institut Dalle Molle D'intelligence Artificielle et Perceptive - Switzerland.

The experiments carried out have led to the conclusions summarized below (see figure 2) :

- the distinction between the oral or nasal feature of a vowel can be reliably made from its duration (nasal vowels are longer than their oral counterparts), measured by our phoneme models (see figure 3) ;
- no significant difference has been measured in the duration of low and high vowels in any position of the syllable in the word ;
- as expected, the position of the vowel in the word highly influences its length (word final vowels tend to be longer because of final lengthening) ;
- the influence of the consonantic context is not reliably measurable, although it can be noted that vowels followed by a voiced consonant are, on average, longer than others.

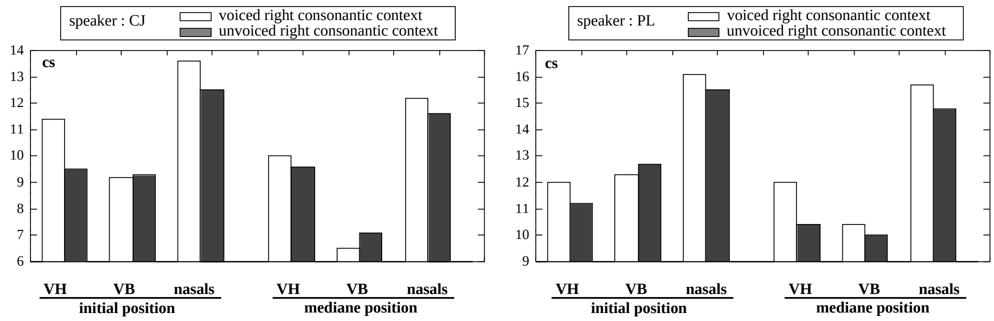

Figure 2: *High (VH), low (VB) and nasal vowel average of duration measures (in cs) in initial and middle position in the FeLex words in voiced and unvoiced right consonantic contexts. Durations were obtained after an alignment process with the non contextual phonemes models described below.*

These experiments lead to the conclusion that it is not reasonable to propose reliable ponderation coefficients allowing to correct the duration variations linked to the microprosodic effects. Furthermore, no index robust enough to be integrated into an ASR system has been found (either for lack of reliability, or simply because they did not improve the SPEX acoustic-phonetic decoding process).

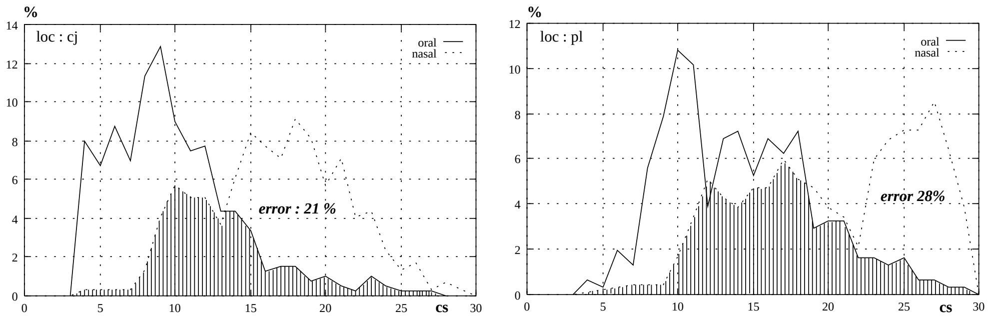

Figure 3: *Distributions of oral and nasal vowels duration measured for 2 speakers (by our Markovian phoneme models) on FeLex (800 words). The x-axis represents the vowel duration (in cs) and the y-axis the percentage of vowels in the Felex corpus with a given duration. The error probability of Bayesian discrimination by means of these distributions is also reported. Similar distributions are observed for other speakers and other corpora.*

2.2 Fundamental frequency

Many algorithms used to compute the fundamental frequency of a speech signal have been described [Rabiner, 78, Hess, 83, Kießling, 92, Ohmura, 94]. After testing some that often fail with irregular speech portions, we decided to develop a relatively simple and classical algorithm which filters the speech signal, computes average magnitude difference functions (AMDF) every 10 ms, identifies the voiced/unvoiced portions of the signal and finds the optimal path between AMDF coefficients (5 candidates per 10 ms maximum).
Particular attention has been paid to the voiced/unvoiced decision which involves not only the usual parameters (zero crossing, spectral energy ratio, etc,), but also the shape of the AMDF curve at a given time [Langlais, 92]. A voiced/unvoiced curve allowing a fine distinction is then provided.
This F0 computing method yields good results in various contexts (different kinds of speakers, telephone speech quality, noisy signals) and preserves microvariations (the smoothing effect can be adjusted).
The systematic examination of the intrinsic and co-intrinsic fundamental frequency variations has provided contrasting results (illustrated in figure 4) leading to the following conclusions :

- high vowels have, on average, a higher fundamental frequency than low ones, except at the end of words where no difference has been measured. Yet, this difference seems too small for a functional discrimination between the two kinds of vowels. Intrinsic values could however be expected to play a larger role on word final vowel than on word other vowels in sentences ;
- the influence of the consonantic context is not obvious from our measures although it can be noticed that, on average, the F0 of the vowels is higher when they follow an unvoiced consonant ; this difference can however not be exploited in a recognition process ;
- in an intervocalic context, the consonants [ℓ], [m] and [n] are not supposed to exhibit local lowering of the F0 curve, unlike other voiced consonants. This index allows an 80% classification rate (see figure 5).

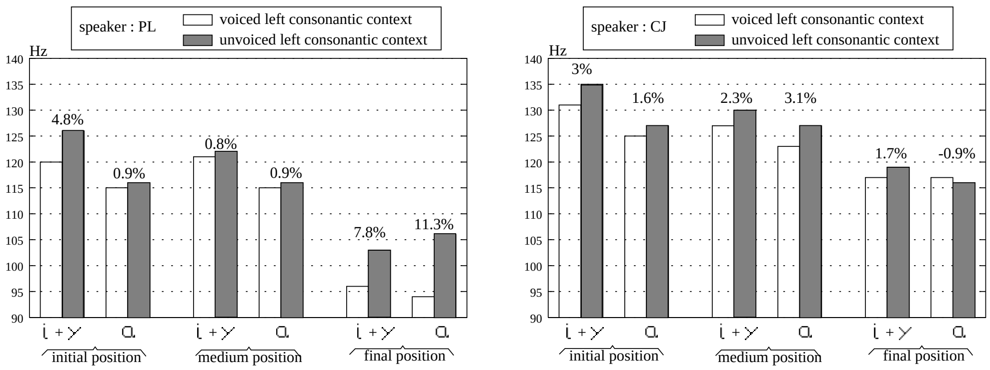

Figure 4: *Average fundamental frequency values of high vowels [i] and [y], and the low vowel [a] in each syllabic position (initial, middle, final) of the Felex words in voiced and unvoiced left consonantic context.*

In addition to all these factors, we have tested the possible contribution of the voiced/unvoiced index to lexical filtering. Three filters were implemented in SPEX and tested on a

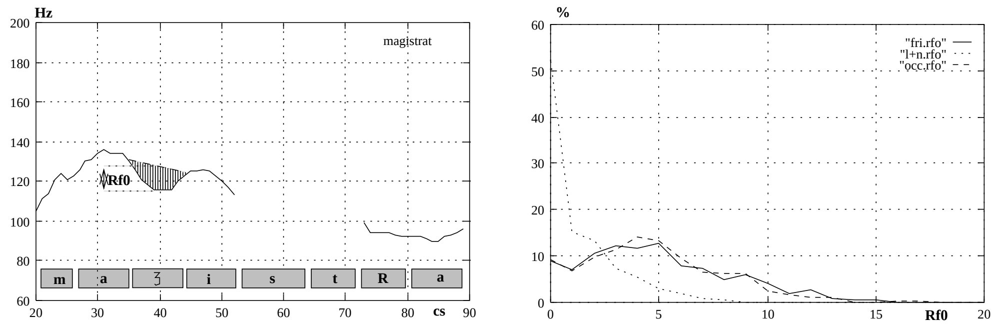

Figure 5: **a)** *Illustration of the F0 curve of the intervocalic voiced consonant [z] in the French word magistrat (magistrate).* **b)** *Distribution of the concavity measure (Rf0) made on intervocalic voiced consonants (fri=fricatives, occ=stops, l+n=[ℓ]+[n]+[m]). Voiced fricatives and voiced stops frequently show a fall-rise pattern, contrary to the [ℓ], [n] and [m] consonants, whose F0 pattern is generally flat.*

500-word corpus, with each word uttered by 5 speakers (4 males, 1 female) in "laboratory" conditions [Béchet, 95] . These 2500 were to be recognized among a lexicon of 20 000 words. The three filters involving the voiced/unvoiced decision mentioned above can be described as follows :

- Before the acoustic-decoding, a first filter allows access only to the lexical units whose voiced/unvoiced theoretical patterns coincide with the effective labelling of the signal. Even if this process requires the use of rules governing the pronunciation variations (assimilation, elision of [ə], etc.), this kind of filtering is easy to carry out and does not depend much on the speaker. The results are quite interesting since a filtering rate of 60% was measured with a 3% error rate (less than 1% with a slight speaker adaptation).
- After the acoustic-phonetic decoding process, SPEX segments the speech signal into vocalic and consonantic macro-classes (vowels, voiced and unvoiced consonants, etc.) which can be composed of several phonemes. According to this principle, the pre-encoded segmentations of the lexicon units allow to align approximately candidate words phonemes on the signal. A simple verification of the matching of the voiced/unvoiced index with the corresponding zones allows the elimination of the improbable hypotheses. This extremely simple filtering is also very efficient since it eliminates 25% of the remaining words with an error rate inferior to 1%.
- At the end of the process, SPEX proposes a set of valued lexical hypotheses (of about 100 words) with their associated temporal alignments. A more precise voiced/unvoiced filtering removes bad candidates. The filtering rate is still interesting : on average, 20% of the words in the set are removed with a relatively small error rate (3%) ; the rank of words not recognized in the first position is improved by 3 positions.

These experiments leads to the conclusion that the voiced/unvoiced feature as it is measured seems to be a promising index for ASR systems. Further experiments need to be carried out to confirm this first observation on a larger speech recognition task (ex : in a speaker independant Markovian continuous speech recognition system).

2.3 Intensity

Intensity is a very often neglected parameter. Several reasons can account for this lack of interest : the partial correlation of its variations with the ones of the F0, the difficulty in evaluating it, its variability depending on sounds and context, and the perceptive normalization problems [Rossi, 71, Rossi, 78]. The analysis of the intrinsic and co-intrinsic intensity of the vowels confirms partially the previous studies. This information allows the distinction only of the sounds well discriminated by other indexes (ex : distinction of the vowels [*a*] and [*i*] at the beginning and at the end of the word with an associated error probability of 0.18 and .08).

3 Macroprosody

This section presents the interest and difficulties of integrating suprasegmental information into ASR systems. An approach is then suggested to address this problem. Finally, the results obtained on two continuous speech corpora are reported.

3.1 Interest and difficulties

Many reasons justify the interest of integrating prosody into speech recognition :

- reasons inherent to the current ASR systems which seem to have reached their limits for the time being. Other sources of knowledge, under-exploited until now, and to which prosody belongs, are therefore attractive solutions to turn to.
- reasons specific to prosody : Di Cristo [Di Cristo, 81] has reviewed the numerous functions which prosody serves in the communication process. We can mention the ones which seem the most promising in speech recognition : the grouping of semantically related words, the detection of sentence modality, the disambiguation, etc.

We might wonder why prosody has not found its place in recent speech recognition systems yet. This is due to difficulties linked on the one hand to the systems themselves and on the other hand to the prosodic analysis. For the former, we can notice the following :

- systems suffer from limitations which make the existing prosodic models unsuitable (gaps of syntactic [Caelen, 89] and even more semantic components [Caelen, 91]),
- fusion of a prosodic component — as far as it is available — with other information sources (acoustic and other) remains a fundamental problem that we cannot get around,

For the latter, we can mention the following :

- automatic prosodic parameter extraction methods are not fully reliable. Furthermore, it has often been said that raw parameters should be corrected by perceptual considerations. Despite studies showing the multi-parametrical aspect of speech perception [Munson, 47, Rossi, 72, Rossi, 76], knowledge on how duration, F0 and intensity are perceived and processed in continuous speech is not advanced enough to be applicable in an ASR system [Vaissière, 88].

- as ASR systems aim at spontaneous speech recognition, and despite the growing number of prosodic studies on the subject [Ward, 91, Shriberg, 92, Blaauw, 94, Nakatani, 94], it is rather complicated to model prosodic knowledge (frequent hesitations, repairs, emphasis, agrammatical sentences, etc.).
- above all, the basic problem of prosodic analysis is that prosodic parameters are governed at the same time by several organizational levels which may become conflicting (syntax, semantic, rhythm, ...) and like Hirst [Hirst, 84, p. 43], we must admit that despite the large number of studies, the problem is far from being solved.

To address this problem (at least partially), the tendency in recent work has been to introduce statistical tool components ; either for Text to Speech applications [Traber, 90, Hirschberg, 94] or for prosodic features recognition applications such as automatic boundary spotting and/or classification [Kompe, 94, Kuhn, 94, Shimodaira, 94, Pagel, 95], recognition of phrase-level intonation contours [Jensen, 94], relationship between prosodic and syntactic structures for sentence disambiguation purposes [Veilleux, 93, Hunt, 94], etc. We support this trend. However, we should bear in mind that statistics is not an end in itself but a mere tool for prosodic analysis [Schneider, 95].

3.2 Bottom-up prosodic labelling

Prosodic labelling consists in reducing all the prosodic parameters to a limited set of labels with a minimum loss of information. This problem has not found a unanimous solution yet [Bruce, 89, Silverman, 92, Hirst, 94, Wightman, 94, Astesano, 95]. Therefore, a fairly classical set of 40 labels is retained in this experiment. These labels (9 duration labels, 9 intensity labels and 22 F0 labels) characterize each vocalic nucleus : maximum and minimum of each parameter for the sentence, emergence of the value of a parameter for a nucleus (n) from the corresponding values on the adjacent segments ($n-2$, $n-1$, $n+1$, $n+2$), several level codings of the parameters from their amplitude measured on the whole sentence, etc.
This simple and fully automatic labelling process (which requires only the speech signal as input) does not include any microprosodic or perceptual corrections, for the reasons mentioned above. It is obvious that a prosodic labelling process is open to critiscism. Therefore, new labels can be added to this process (as far as they can be automatically computed) or, on the contrary, some of them can be suppressed. An example of prosodic lattice is given in figure 6.

3.3 General overview of the system

The existing systems have all contributed to showing that formalized prosodic rules are not very efficient in speech recognition and generate many exceptions [Vaissière, 76, Carbonell, 82, Méloni, 82, Mertens, 87, Nasri, 90, Bonin, 90].
Nevertheless, we formulate the hypothesis ($\mathcal{H}$) that the distribution of prosodic configurations on the whole sentence is not random, but on the contrary regular enough to permit structural (syntactic, rhythmical, semantic, ...) predictions in an automatic way. This is a point we wanted to make with our system (called Pro&Stat) which has two major goals :

- to allow the user to rapidly and automatically obtain correlations between prosodic labels and different organization levels of the message ;

- to automatically propose for a given unknown sentence a set of valued structural hypotheses which can be used in speech recognition ;

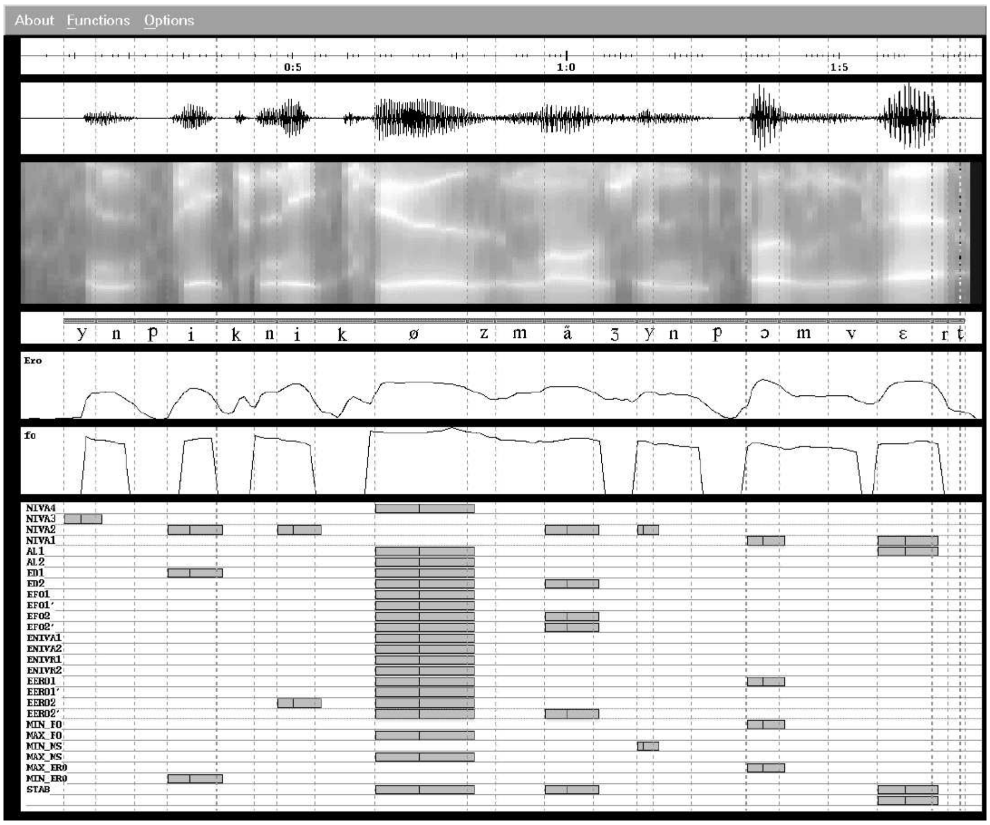

Figure 6: *Example of a prosodic lattice (bottom-up frame) obtained for an utterance of the sentence :* une pique-niqueuse mange une pomme verte *(A picnicker is eating a green apple).*

The main characteristic of the system is that it does not have any a priori either on the hypothetical hierarchy of the different constraints governing the prosodic parameters, or on particular prosodic entities such as stress for which we have found no robust acoustic characterization. The basic principle of Pro&Stat is a data-driven study of correlations between linguistic levels and prosodic labels.
The scope of this experiment is reduced to the automatic study of rhythmical and syntactic interactions with the prosodic labels distribution.

3.3.1 Learning phase

Prosodic labels and a syntactic tree (whose leaves are aligned on the speech signal) are provided (automatically or not) for each sentence of a learning corpus. The formalism used for the tree description respects the following rules (α and β belonging to the user's set of symbols :) :

- $\star$ $\alpha(\beta, n)$ means α is a group of n vowels described by β,
- $\star$ $\alpha(\epsilon, n)$ means α is a leaf of n vowels,
- $\star$ $\alpha(\beta, \varepsilon)$ means α is a group described by β with no mention of its number of vowels,
- $\star$ $\alpha.\beta$ means α and β,

An example of such a structure is given in figure 7. The learning stage consists in plotting a graph which can be described as follows :

- each node (p) contains the prosodic characterization of a syntactic-rhythmical structure ($\mathcal{S}_p$),
- each arc (a_{ij}) stand for a syntactic or rhythmical constraint refining the structure described by node i .

Arcs are defined from the learning sentence structures by applying, level by level, one of the following rules (see figure 7 for an example) :

$\alpha(\beta, n)$	$\rightarrow \alpha(\epsilon, n)$	an α group of n vowels described by β is also an α group with n vowels
	$\rightarrow \mathcal{G}(\epsilon, n)$	an α group of n vowels described by β is also a group of n vowels
	$\rightarrow \alpha(\epsilon, \varepsilon)$	an α group of n vowels is also an α group
	$\rightarrow \mathcal{G}(\epsilon, \varepsilon)$	an α group of n vowels is also a group

where $\mathcal{G}$ is a symbol for any type of group.

Each node keeps count of the labels located at the beginning or at the end of the associated structure leaves, in such a way that each syntactic-rhythmical structure is prosodically described by means of a (normalized) matrix with n lines (n is the number of the different labels – 40 in this study) and k columns (k is maximized by the number of leaves of the described structure $\times$ 2).
After the learning stage — which is fully described in [Langlais, 95, pp. 134 to 138] — a user can ask the system to provide him with meeting points between syntax, rhythm and prosody, and can also ask for a parametric contour matching a given linguistic organization.

3.3.2 Structural prediction phase

This stage consists in putting forward syntactic-rhythmical hypotheses from the learning graph (G) and the prosodic labelling of the speech signal to decode (considered as an $n \times l_o$ matrix M_o, l_o being the assumed number of vowels). This process involves the parser briefly described below (see [Langlais, 95, pp. 141 to 144] for further details).
Given the following notations : p is a node belonging to G, describing a structure ($\mathcal{S}_p$) of f leaves ; it is prosodically described by an $n \times l_p$ matrix (with $f < l_p < 2 \times f$). The parser requires two basic steps :

M_o reduction from $\mathcal{S}_p$: consisting in finding a set of f couples of M_o columns (d_i, f_i) corresponding to the possible limits (in terms of vowels) of the f constituents. This process verifies the constraints on the number of vowels of each $\mathcal{S}_p$ leaf as well as the following conditions : for each $i \in [1, f], d_i = f_i + 1, f_i \geq d_i, d_1 = 1 and f_{l_p} = l_o$. The resulting matrices are noted M_{o_p}.

similarity measurement : between two matrices M_{o_p} and M_p (which have the same dimension) by computing $d(M_{o_p}, M_p)$:

$$\text{with} \left\{ \begin{array}{lll} d(M_{o_p}, M_p) & = & \frac{\sum_{i=1}^{n} \left(\alpha_i \sum_{j=1}^{l_p} \delta_{ij} \right)}{n} \\ \text{et } \delta_{ij} & = & \left\{ \begin{array}{lll} M_p(i,j) & if & M_{o_p}(i,j) = 1 \\ 0 & else & \end{array} \right. \\ \text{and } \alpha_i & & \text{weight of the } i^{th} \text{ label} \end{array} \right.$$

The valuation of an hypothesis (any node of the learning graph) is achieved by keeping the best score path from the root of G to the considered node. The scoring of a particular path is the average of the similarity measurement of M_{o_q} and M_q matrices for each node q of the path.

a) $PH($ $SS(GN(Art(\epsilon,1).NC(\epsilon,3),4),4).$
$SV(VB(\epsilon,1).GN(ART(\epsilon,1).NC(\epsilon,1).ADJ(\epsilon,1),3),5)$
$,9)$

b) $PH(SS(\epsilon,4).SV(\epsilon,5),9),$
$PH(\mathcal{G}(\epsilon,4).\mathcal{G}(\epsilon,5),9),$
$PH(SS(\epsilon,\varepsilon).SV(\epsilon,\varepsilon),9),$
$PH(\mathcal{G}(\epsilon,\varepsilon).\mathcal{G}(\epsilon,\varepsilon),9)$

Figure 7: **a)** *Ex : syntactic-rhythmical structure given for the utterance of the sentence of figure 6 (une pique-niqueuse mange une pomme verte) with the set of user's symbols : PH sentence, SS subject group, GN noun group, ART article, NC common noun, SV verb group, VB verb, ADJ adjective.* **b)** *nodes created by the application of the production rules to level 2 of the structure described in a).*

3.4 Results

Our main objective is to verify the above mentioned work hypothesis ($\mathcal{H}$). Consequently, no particular attention has been paid to improving the results (weighting of the prosodic labels, more suitable syntax, ...). Possible improvements are discussed in sub-section 3.5.

3.4.1 Isolated sentences

A corpus of 500 isolated sentences (newspaper excerpts), made of 80 different sentences of simple syntactic structure from 4 to 17 vowels (see examples in figure 8) uttered by 50 speakers via a noisy telephone line, has been used for learning purposes. The syntactic trees have been hand-labelled and the prosodic lattices where automatically computed.

⋆ Le tout apparaît dans un bilan clair.
(Everything appears in a clear conclusion.)
⋆ Cette révélation tardive souligne la gravité du malaise.
(This late revelation highlights the seriousness of the situation.)

Figure 8: *Example of learning sentences.*

We ask our system to make full syntactic-rhythmical hypotheses of unknown sentences (full syntactic tree, as well as the number of vowels for each leaf) only from the prosodic lattice. The average number of possible answers for a given sentence is 15.

It has shown interesting capacities which allow its use for sentence recognition. On the learning corpus, more than 90% of the hypotheses have been classified in first position (see figure 9a). In a test corpus of 300 sentences (most of them being repetitions of the original 80 sentences of the learning corpus, not necessarily by the same speakers), the first hypothesis assigns almost 60% of sentences to the right structure (see figure 9b).

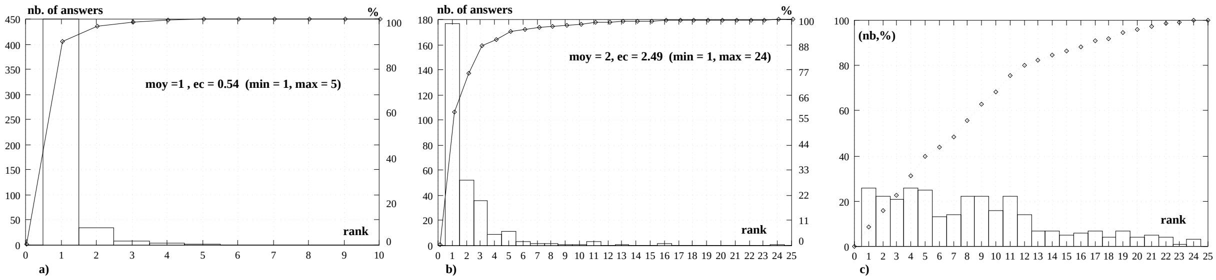

Figure 9: *Isolated sentences : the x-axis represents the rank of the hypothesis correctly formulated by the system; the y-axis indicates, on the left, the number of hypotheses and, on the right, the percentage of correct hypotheses formulated at a given rank.* **a)** *predictions on the learning corpus (500 sentences).* **b)** *predictions on the testing corpus (300 sentences).* **c)** *predictions on the testing corpus by means of random scoring.*

3.4.2 Decimal numbers

A corpus of 500 decimal numbers (see examples in figure 10) uttered by 50 speakers via the same telephone line has been used for learning purposes (prosodic lattice and syntactic trees were obtained automatically).

⋆ Treize mille deux cent quarante deux virgule dix sept.
(Thirteen thousand, two hundred forty two point seventeen.)
⋆ Quatre mille virgule huit.
(For thousand point eight.)

Figure 10: *Example of decimal numbers from the learning corpus.*

Here again, the system has been able to match the prosodic lattices of the learning numbers to the correct syntactic-rhythmical structures in more than 80% of the cases, with an average of 15 possible answers (see figure 11a). On a test corpus of 148 numbers, the system has shown a prediction rate of slightly less than 50% (see figure 11b). The analysis of the non-first hypotheses has shown that the word "virgule" (point) was very often well located (see figure 11c). We were not able to reach this score by means of local specific rules, which confirms the importance of the prosodic information taken from the whole sentence.

3.5 Further developments

These results validate Pro&Stat and allow us to conclude that the prosodic organization of the sentence as an entity does exist and can be handled through an automatic approach. Whatever the case, numerous problems and difficulties must be taken into account in order to deal with prosody. Some of the important points are mentioned below :

- testing other prosodic labels,
- weighting each label and removing redundancy,
- associating the labels in a more sophisticated way in order to define larger and more structured prosodic units (especially on the whole syntagm)
- applying the system to more specific tasks such as recognition of the speaker's language, disambiguation, etc,
- implementing a recognition system — even limited — which achieves optimal fusion between prosodic, acoustic-phonetic and linguistic knowledge,
- broadening the study to richer syntactic structures and to other organizational levels
- widening the range of speech types (spontaneous speech, dialogue, etc.)
- increasing the size of the corpus for learning and testing, etc.

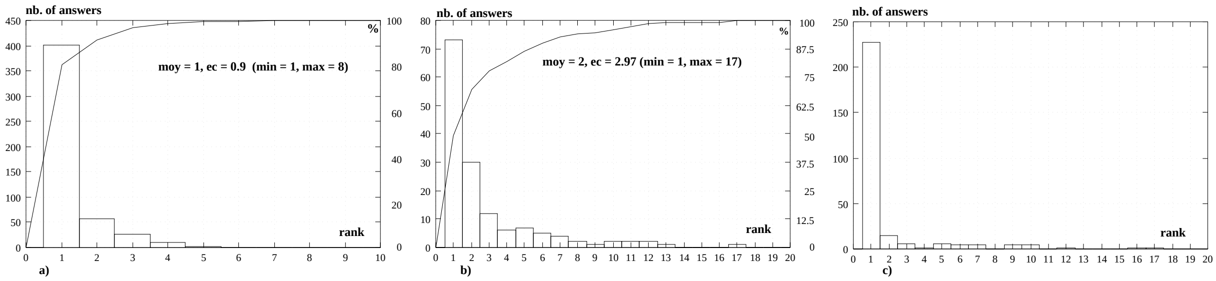

Figure 11: *Decimal numbers :* **a)** *predictions on the learning corpus (500 numbers).* **b)** *predictions on a sub-set of the testing corpus (the numbers whose syntactic-rhythmical structure has been encountered at least once during the learning process, i.e. 148 numbers).* **c)** *predictions of the place of the word "virgule" (point) in the numbers of the testing corpus (280 numbers).*

4 Conclusion

In light of our experiments, we discussed the possible integration of prosody into a speech recognition frame.
We first studied some possible uses of microprosodic fluctuations in an acoustic-decoding process and concluded that such indexes are weak. We tested an interesting voiced/unvoiced decision in an isolated word recognition system and further experiments should be carried out to analyse its behavior in continuous speech recognition systems.
Secondly, we proposed a system which allows to verify that prosodic interactions between syntax and rhythm could be approached in an automatic way. Some efforts remain to be made in order to put this under-exploited knowledge to full use, all the more so since the opening of recognition systems to spontaneous speech witch will give much more importance to prosodic phenomena.
Moreover, prosody implies many other phenomena than the ones linked to the nature of sounds and to the linguistic structuration of sentences. Therefore, it is essential to widen our research in order to extract other information such as speaker identification, emotions, variations of style, speech pathologies, social and dialectal variations, language learning assistance, etc.

References

[Astesano, 95] C. Astesano, A. Di Cristo, and D. Hirst. Discourse-based empirical evidence for a multi-class accent system in french. In *XIIIth International Congress of Phonetic Sciences*, volume 4, pages 630–633, Stockholm, 13–19 August 1995.

[Béchet, 94] F. Béchet. *Système de traitement de connaissances phonétiques et lexicales : Application à la reconnaissance de mots isolés sur de grands vocabulaires et à la recherche de mots cibles dans un discours continu.* PhD thesis, Université d'Avignon et des Pays de Vaucluse, janvier 1994.

[Béchet, 95] F. Béchet, P. Langlais, and H. Méloni. Lexical filtering by means of prosodic information. In *International Congress of Phonetic Sciences*, volume 4, pages 336–339, Stockholm, Sweden, August 13–19 1995.

[Blaauw, 94] E. Blaauw. The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech. In Elsevier Science B.V., editor, *Speech Communication*, volume 14, pages 359–375, 1994.

[Bonin, 90] J.-J. Bonin and J.-M. Pierrel. Fréquence fondamentale et durée pour la détection de frontières syntagmatiques en parole continue. In *XVIIIème Journées d'Étude sur la Parole*, pages 21–25, Montréal, 28-31 mai 1990.

[Bruce, 89] G. Bruce. Report from the IPA working group on suprasegmental categories. Working papers, 35, Lund University Department of Linguistics and Phonetics, 1989.

[Caelen, 89] G. Caelen-Haumont. Une représentation syntaxique adaptée à la prosodie. In *Journal d'Acoustique*, volume 2, pages 137–146, 1989.

[Caelen, 91] G. Caelen-Haumont. Stratégies des locuteurs en réponse à des consignes de lecture d'un texte : analyse des interactions entre modèles syntaxiques, sémantiques, pragmatique et paramètres prosodiques. Thèse de doctorat d'état - spécialité es lettres, Institut de la Communication Parlée, Grenoble, 29 novembre 1991. Soutenue à Aix-en-Provence.

[Campbell, 92] W. Campbell. Prosodic encoding of English speech. In *International Conference on Spoken Language Processing*, pages 663–666, Banff, Canada, 1992.

[Carbonell, 82] N. Carbonell, J.-P. Haton, F. Lonchamp, and J.-M. Pierrel. Élaboration expérimentale d'indices prosodiques pour la reconnaissance ; application à l'analyse syntaxico-sémantique dans le système Myrtille II. Séminaire Prosodie et Reconnaissance d'Aix-en-Provence, Octobre 1982.

[Chollet, 95] G. Chollet, J.-L. Cochard, P. Langlais, and R. van Kommer. Swiss-french polyphone: a telephone speech database to develop interactive voice servers. In *Linguistic Databases*, Gröningen, 1995.

[Di Cristo, 75] A. Di Cristo. *Soixante et dix Ans de Recherches en Prosodie. Bibliographie alphabétique, thématique et chronologique.* Université de Provence, 1975.

[Di Cristo, 78] A. Di Cristo. *De la microprosodie à l'intonosyntaxe.* PhD thesis, Université de Provence, 1978.

[Di Cristo, 81] A. Di Cristo. Aspects phonétiques et phonologiques des éléments prosodiques. *Modèles linguistiques*, Tome III(Fascicule 2):24–83, 1981. Presse universitaire de Lille.

[Dumouchel, 94] P. Dumouchel. Suprasegmental features and continuous speech recognition. In *ICASSP*, volume II, pages 177–180, 1994.

[Hess, 83] W. Hess. *Pitch determination of speech signals*. Springer-Verlag, 1983.

[Hirschberg, 94] J. Hirschberg and P. Prieto. Training intonational phrasing rules automatically for english and spanish text-to-speech. In *Second ESCA/IEEE workshop on Speech Synthesis*, pages 159–162, New York, September 12-15 1994.

[Hirst, 84] D. Hirst. Prosodie et structures de données en phonologie. In Hermann, editor, *Formes sonores du langage. Structure des représentations en phonologie*, pages 43–63. F. Dell and D. Hirst and J.R. Vergnaud, 1984.

[Hirst, 94] D. Hirst, N. Ide, and J. Véronis. Coding fundamental frequency patterns for multi-lingual synthesis with INTSINT in the MULTEXT project. In *Second ESCA/IEEE workshop on Speech Synthesis*, pages 77–80, New York, September 12-15 1994.

[Hunt, 94] A. Hunt. A generalised model for utilising prosodic information in continuous speech recognition. In *ICASSP*, volume II, pages 169–172, Adelaïde (Australia), 1994.

[Jensen, 94] U. Jensen, R. Moore, P. Dalsgaard, and B. Lindberg. Modelling intonation contours at the phrase level using continuous density hidden markov models. In *Computer Speech and Language*, volume 8, pages 247–260, 1994.

[Kießling, 92] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. DP-based determination of f0 contours from speech signals. In *ICASSP*, volume II, pages 17–20, San Francisco, 1992.

[Klatt, 75] D. Klatt and W. Cooper. *Perception of segment duration in sentence contexts*, pages 69–86. Springer Verlag, Berlin, communication & cybernetics edition, 1975.

[Kompe, 94] R. Kompe, A. Batliner, A. Kießling, U. Kilian, E. Niemann, H.and Nöth, and P. Regel-Brietzmann. Automatic classification of prosodically marked phrase boundaries in german. In *ICASSP*, volume II, pages 173–176, Adelaïde (Australia), 1994.

[Kuhn, 94] R. Kuhn, A. Lazarides, Y. Normandin, J. Brousseau, and E. Nöth. Applications of decision tree methodology in speech recognition and understanding. In *CRIM/FORWISS Workshop*, pages 220–232, September 1994.

[Langlais, 92] P. Langlais, H. Méloni, and J. Vaissière. Étiquetage prosodique ascendant. In *19èmes Journées d'Étude sur la Parole*, Bruxelles, 1992.

[Langlais, 95] P. Langlais. *Utilisation de la prosodie en reconnaissance de la parole*. PhD thesis, Université d'Avignon et des Pays de Vaucluse, 1995.

[Lea, 80] W. A. Lea. Prosodic aids to speech recognition. In W. Lea, editor, *Trends in Speech Recognition*, chapter 8, pages 166–205. Prentice Hall, New Jersey, 1980.

[Lehiste, 76] I. Lehiste. *Contemporary issues in experimental phonetics*, chapter 7: Suprasegmental features of Speech, pages 225–239. Academic Press, N. Lass edition, 1976.

[Méloni, 82] H. Méloni and J. Guizol. Utilisation de paramètres prosodiques dans un système de reconnaissance automatique de la parole continue. Séminaire Prosodie et Reconnaissance d'Aix-en-Provence, Octobre 1982.

[Mertens, 87] P. Mertens. *L'intonation du français. De la description linguistique à la reconnaissance automatique.* PhD thesis, Université de Leuven, 1987.

[Mertens, 90] P. Mertens. *Le français parlé. Études grammaticales*, chapter IV, pages 159–176. Sciences du langage, CNRS edition, 1990.

[Munson, 47] W. Munson. The growth of auditory sensation. *J.A.S.A.*, 19:584–591, 1947.

[Nakatani, 94] C. H. Nakatani and J. Hirschberg. A corpus-based study of repair cues in spontaneous speech. *Acoustical Society of America*, 95(3):1603–1616, March 1994.

[Nasri, 90] M. Nasri, G. Caelen-Haumont, and J. Caelen. Utilisation de règles prosodiques en reconnaissance de la parole. In *XVIIIèmes Journées d'Étude sur la Parole*, Montréal, 28-31 mai 1990.

[Ohmura, 94] H. Ohmura. Fine pitch contour extraction by voice fundamental wave filtering method. In *ICASSP*, volume II, pages 189–192, Adelaïde (Australia), 1994.

[Pagel, 95] V. Pagel, N. Carbonell, and J. Vaissière. Spotting prosodic boundaries in continuous speech in French. In *XIIIth International Congress of Phonetic Sciences*, volume 4, pages 308–311, Stockholm, 13–19 August 1995.

[Pérennou, 92] G. Pérennou, M. de Calmès, I. Ferrané , and J. Pécatte. Le projet BDLEX de bases de données lexicales du français écrit et parlé . In *Actes du séminaire lexique*, pages 41–56, Toulouse, 1992. IRIT-UPS.

[Rabiner, 78] L. Rabiner and R. Schafer. Digital processing of speech signal. In *Signal Processing Series*. Prentice-Hall, Alan V. Oppenheim series edition, 1978.

[Rossi, 71] M. Rossi. L'intensité spécifique des voyelles. In A. Rigault and R. Charbonneau, editors, *Seventh International Congress of Phonetic Sciences*, pages 574–588. Mouton, 22-28 august 1971.

[Rossi, 72] M. Rossi. Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole. *Phonetica*, III:229–237, 1972.

[Rossi, 76] M. Rossi. Interaction des glissements d'intensité et des glissements de fréquence. In *XIVth International Conference on Acoustics*, High Tatra, 1976.

[Rossi, 78] M. Rossi. The perception of non repetitive intensity glides on vowels. *Journal of Phonetics*, 6:9–18, 1978.

[Rossi, 81] M. Rossi, A. Di Cristo, P. Martin, and Y. Nishinuma. *L'intonation de l'acoustique à la sémantique.* Klincksieck, Paris, 1981.

[Rossi, 85] M. Rossi. L'intonation et l'organisation de l'énoncé. *Phonetica*, 42:135–153, 1985.

[Rossi, 93] M. Rossi. A model for predicting the prosody of spontaneous speech (PPSS model). In *Speech Communication*, volume 13, pages 87–107, North-Holland, 1993. Elsevier Science Publishers.

[Schneider, 95] J.-J. Schneider. *Un système d'apprentissage numérique et symbolique pour la formalisation de connaissances prosodiques.* PhD thesis, Université d'Avignon et des Pays de Vaucluse, 1995.

[Shimodaira, 94] H. Shimodaira and M. Nakai. Prosodic phrase segmentation by pitch pattern clustering. In *ICASSP*, volume II, pages 185–188, Adelaïde (Australia), 1994.

[Shriberg, 92] E. Shriberg, J. Bear, and J. Dowding. Automatic detection and correction of repairs in human-computer dialog. In Harriman, editor, *Speech and Natural Language Workshop*, pages 419–424, 1992.

[Silverman, 92] K. Silverman, M. Beckman, J. Pitrelli, C. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. Tobi : a standard for labeling english prosody. In *ICSLP*, volume 2, pages 867–870, Banff, Alberta, Canada, 1992.

[Traber, 90] C. Traber. F0 generation with a database of natural f0 patterns and with a neural network. In *ESCA workshop on Speech Synthesis*, pages 141–144, Autrans, Septembre 1990.

[Vaissière, 76] J. Vaissière. *Automatic procedure for segmenting continuous speech into prosodic words, in French*, volume 2, pages 193–208. Recherches Acoustiques, Centre National d'études des Télécommunications edition, 1976.

[Vaissière, 82] J. Vaissière. *A suprasegmental component in a french speech recognition system : Reducing the number of lexical hypotheses and detecting the main boundary*, volume VII, pages 109–125. Recherches Acoustiques, Centre National d'Études des Télécommunications edition, 1982.

[Vaissière, 84] J. Vaissière. Proseidon : Automatic detection of prosody clues in continuous speech. In *XIIIèmes Journées d'Étude sur la Parole*, Bruxelles, 1984.

[Vaissière, 88] J. Vaissière. The use of prosodic parameters in automatic speech recognition. In H. Niemann et al., editor, *Recent Advances in Speech Understanding and Dialog Systems*, volume F46. NATO ASI Series, Springer-Verlag Berlin Heidelberg, 1988.

[Veilleux, 93] N. Veilleux and M. Ostendorf. Probabilistic parse scoring with prosodic information. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 51–54, Minneapolis, Minnesota, 27-30 April 1993.

[Ward, 91] W. Ward. Understanding spontaneous speech : the phoenix system. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 365–367, Toronto, 1991.

[Wightman, 94] C. W. Wightman and M. Ostendorf. Automatic labeling of prosodic patterns. In *IEEE transactions on speech and audio processing*, volume II, pages 469–481, 1994.

[Young, 90] S. Young and P. Woodland. *HTK Manual version 1.4.* Cambridge University Engineering Department, 1990.