Rhythmic patterns and their automatic retrieval in spontaneous French

Katarina Bartkova

France Telecom R&D
2, avenue P. Marzin, 22307, Lannion, France
katarina.bartkova@orange-ftgroup.com

Abstract

An auditory analysis of a speech data base in French highlighted the use of rhythmic patterns by speakers. The rhythmic patterns are built by employing successive prosodic groups of the same length expressed in a number of syllables. These rhythmic patterns were not present to the same extent in the elocution of every speaker, but were speaker dependent. Although the auditory detection of the prosodic units in the speech signal is based on a whole group of acoustic parameters, an attempt is made here to retrieve the auditory detected prosodic units automatically. Therefore, one of the goals of this study is to evaluate and approximate the group of acoustic parameters used during auditory segmentation by three parameters such as the vowel duration, vowel energy and the slope of the F0 (measured on the last vowel of the prosodic unit), as they are easy to calculate and are considered as playing a prominent role in the prosodic demarcation of speech.

1. Introduction

The rhythm has a biological and a cognitive origin and influences our production and perception of the speech. But as it is claimed in [6], the rhythm is before all a perceptive construction. In speech, the rhythm is the ability of the speaker to structure the speech during its production. More than a form or a structure, according to [10], the rhythm is a capacity of our cognitive system to "give a form" to the production and the perception of the speech. The perception of the rhythm of the speech supposes a linguistic competence. Though the metric and the rhythm are often considered as synonyms, for [4] the metric implies a more constraint and rigid organization than the rhythm. However, the spontaneous speech cannot cope with rhythm constraints as strict as the poetry, because spontaneous speech must deal with lexical units, syntactic structures, speech rate variation, style variations … There exists a distinction between prepared and spontaneous speech. According to [7], prepared speech favours more metric organisations, whereas the spontaneous character of the speech favours the dynamic organisation, aiming at communicative exchanges.

The exact moments for producing phonetically important elements might obey some larger and possibly context- and language-independent logic. In [13] was shown that speakers appear to structure their utterances as a harmonic fraction of time available to produce sequences of whole phrases. It was also shown [3] that when speakers repeat a short piece of text many times, they exhibit a strong preference for locating prominent (e.g. stressed) syllable onsets at simple harmonic fractions of the repetition cycle.

Stressed syllables constitute a major element in metric organisation: the stressed syllable is considered as a strong beat while the unstressed syllable as a weak beat. The acoustic component of the rhythm in speech is expressed not only by the timing, but also by each element, which participates in the rhythm structuring, creating contrasts expressed by the variation of the prosodic parameters.

A rhythmic based binary classification is made among languages sorting them into stressed- and syllable-timed groups [11]. However, the evaluation of this typology by physical parameters did not lead to straightforward results. In [9] is given an extensive review of the issue of isochrony and the author came to the conclusion that there were no direct acoustic correlates of rhythmicity. According to the author rhythmicity was a perceptual gestalt that emerges from the combination and complex interaction of a great number of acoustic and/or motor parameters. This view, supported by a number of further studies, has essentially formed the consensus for spontaneously produced speech.

A similar idea is supported by the theory of «Perceptif Centre» (P-Center) which claims that in the perception of the rhythm, our internal clock plays a major role [12]. According to this theory, listeners tend to segment in a quasi-periodic way a speech signal, even if the acoustic patterns present in the signal do not reflect this perceived rhythm. However, some rhythmic patterns are still identifiable in speech and they reflect, at the same time, universal constraints and constraints that are language dependent.

The goal of the present study is to analyse the rhythmical organisation of spontaneous speech in French. The regularity of occurrences of prosodic groups having the same length creates a rhythmic pattern of the sentence. Here, the rhythmic patterns are the repeated production of prosodic units having the same length expressed in a number of syllables. Prosodic units are first detected on a speech corpus in an auditory way (i.e. human detection). The physical parameters that play a major role in the perception of rhythmic patterns and are easily accessible from the speech signal, such as vowel duration, vowel energy and the F0 slope, are then measured. Subsequently, an attempt is made to retrieve the manually segmented prosodic units automatically using the measured prosodic parameters.

2. Data base analysis

Our data base contains spontaneous speech in French uttered by 33 speakers. Each speaker uttered several messages of different length and content. The mean length of the messages was 32 words (ranging from 6 to 77 words). The task consisted in leaving a vocal message uttered spontaneously. The speakers were aware that they were recorded, however their messages kept a fully spontaneous character. The content

of messages was hand-transcribed and a first automatic segmentation using forced alignment was carried out. This automatic segmentation was then hand-checked in order to correct major errors of alignment (such as short pause insertions, omissions or insertions of schwas…). As the automatic phonetic transcription contained pronunciation variants, the automatic segmentation turned out to be correct each time the transcription of the messages was free of errors.

2.1. Auditory analysis of the corpus

Once the segmentation carried out, several prosodic parameters were measured on the speech signal, among others the F0 slope on the last vowel of the prosodic unit, the length of the last vowel, the length of the syllables… The mean duration of vowels was calculated for each message and was consequently used to normalise the vowel duration of the message.

The listening of the data allowed its segmentation into prosodic groups. 1207 prosodic groups were auditorily detected in the whole speech corpus and their lengths were expressed in numbers of syllables. As illustrated in Figure 1, the length of most of the prosodic groups was around 3 syllables. In fact, a preference of speakers can be observed to structure the speech into short elements: the number of prosodic groups having a length of 6 syllables, or more, decreases very clearly.

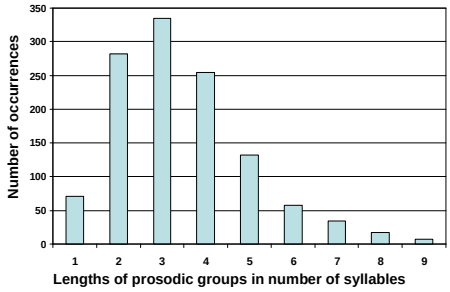


Figure 1: *Distribution of prosodic groups according to their lengths expressed in numbers of syllables.*

The figures about prosodic group lengths, found in our study, are comparable to those published by other researchers. Thus, according to [5], [8] and [16], the interval between 2 stresses has a mean variation ranging from 2 to 5 syllables.

2.2. Distribution of the prosodic groups

Although all speakers of the data base used rhythmic patterns, the rhythmic organisation of the speech could be regarded as one of the characteristics of the elocution of the speaker. In fact, it was observed in our data that some speakers used rhythmic patterns more often than others, i.e., they produced prosodic groups having the same number of syllables more easily. Figure 2 contains cumulative histogram of speakers that produced rhythmic patterns (several successive prosodic groups having the same number of syllables) in their messages. It appears that 50% of speakers produced at least 40 % of their prosodic groups with a rhythmic pattern, and more than 70% of speakers produced at least 30% of their prosodic groups as rhythmic patterns.

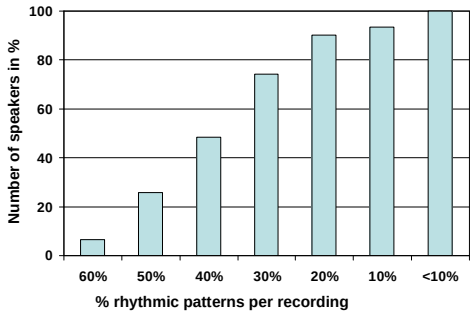


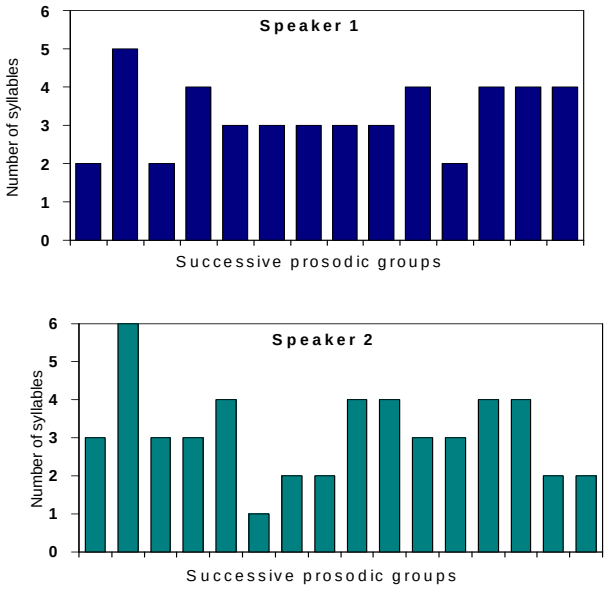
Figure 2: *Cumulative histogram of rhythmic patterns used by the speakers.*

The validity of the use of the syllabic metric to express the length of prosodic groups can be justified by the measure of the vocalic duration values. In fact, the mean duration of a non-final (therefore unstressed) vowel in prosodic groups showed a great homogeneity, no matter how long the prosodic group was: the vowel duration ranged from 60 to 72 ms with a standard deviation of 20 ms. Even the mean duration of the last vowel of the prosodic group (stressed vowel) turned out to be uninfluenced by the length of the prosodic group. Its duration ranged from 130 ms to 160 ms with a standard deviation of 80 ms.

2.3. Detailed analysis of rhythmic patterns

In this paragraph some examples are given of a detailed analysis of rhythmic patterns for some speakers. As already mentioned in the previous paragraph, for certain speakers, rhythmic patterns occurred frequently and they could include a large part of the sentence while for other speakers, very few rhythmic patterns were found.

Figure 3 gives examples of rhythmic patterns for 4 speakers as they were found in 1 message per speaker. The succession of prosodic groups is represented on horizontal axis and the length of the prosodic groups (expressed in numbers of syllable) is represented on vertical axis.



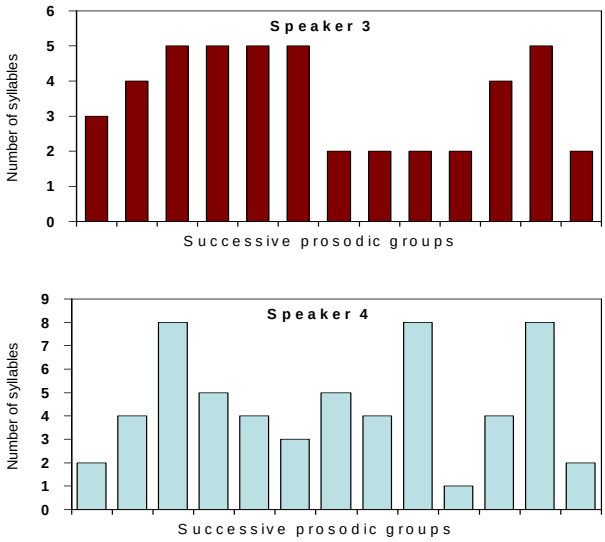


Figure 3: *4 examples of the succession of prosodic groups expressed in numbers of syllables*

Figure 3 shows a high tendency toward eurhythmicity for speaker 1 and for speaker 3. In fact, in case of speaker 1, the prosodic groups, with a length of 3 syllables, are repeated 4 times in a row and thus they create the same rhythmic pattern. The same speaker finishes his message with a large rhythmic pattern, uttering, 3 times, prosodic groups of 4 syllables. Speaker 3 also demonstrates a high eurhythmic tendency. In fact, this speaker creates a first rhythmic pattern uttering four successive prosodic groups of a length of 5 syllables, followed by 4 successive prosodic groups of a length of 2 syllables. Speaker 2 shows a trend toward simple rhythmic patterns as he privileges one simple repetition of prosodic groups of same length. As far as the 4th speaker is concerned, he uses no rhythmic pattern at all; in fact, no repetition of prosodic units of same length was observed in his pronunciation.

3. Automatic retrieval of prosodic boundaries

It seemed challenging to retrieve automatically the prosodic groups found using the auditory segmentation. This task seems to be quite complex if we admit that the detection of the rhythm of speech by listeners is based on a whole range of acoustic and articulatory parameters. As the opinion about pertinent parameters of the speech rhythm diverges in the literature, we chose to use the 3 main prosodic parameters easily accessible from the speech signal, which are, vowel duration, F0 slope and the vowel energy.

3.1. Vowel duration modelling

In the modelling of the vowel duration a distinction is made between vowels occurring in three positions: in stressed position (position found during the auditory checking), in the last syllables of a word in unstressed position and in non-final (unstressed) syllables.

These distinctions were motivated by the results obtained in [2], which showed on 2 data bases of spontaneous French, a significant lengthening of the vocal duration observed in final position of prosodic groups. However, a lengthening, though more moderated, was also found in the last syllables of words occurring out of prosodic boundaries.

As our speech data contains spontaneous speech, we encountered spontaneous speech phenomenon such as filled pauses, speech repairs and word repetitions. It appeared wise to consider filled pauses (generally with vocalic timber of /ø/) as an extra prosodic category as its duration is very long and its F0 is mostly flat [1]. The modelling of the vocal duration was carried out using discrete models and vowel durations were normalized with the mean vowel duration calculated for each recording. The model was trained on 1 message per speaker (about 33 messages in all) and tested on the remaining data (57 messages).

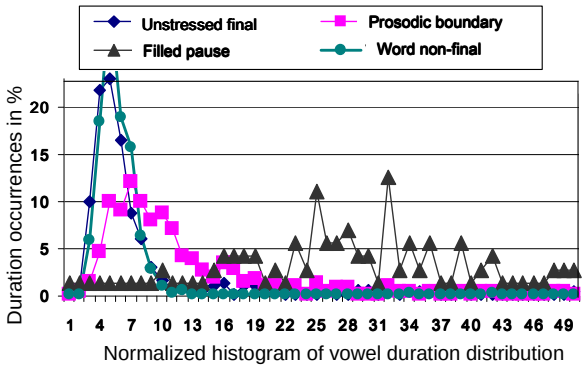


Figure 5: *Normalised histograms of the vowel duration measured in 4 positions*

Figure 5 represents a normalized histogram which captures the distribution of vowel durations during the training of the model. It appears clearly from the figure, that the distinction between positions "**Word non-final**" and "**Unstressed Word final**" becomes impossible, therefore these two categories were merged together. On the other hand, the histogram predicts a good detection of filled pauses (curve of filled pauses clearly separated from the other curves) and a relatively good detection of prosodic boundaries (curve partly covered with the curve representing non-final and unstressed final vowels).

3.2. Test of the duration model

The test of the model was carried out on 57 recordings kept apart for this task. The position of the current vowel was predicted by the model (normalized histograms) corresponding to the greatest likelihood for the observed duration. The results are the following:

Table I : *Correct detection rate obtained by the duration model*

Unstressed vowel	Prosodic boundary	Filled pause
86 %	*67 %*	*98 %*

As foreseen, the model gives good detection for filled pauses and for unstressed vowels (non-final position or final unstressed position). But the detection of prosodic boundaries using only vowel duration is not completely satisfactory. A high confusion (33%) remains between the prosodic boundaries and the unstressed vowels. This proves that though

the role of vowel duration in prosodic parsing is important, it does not suffice alone to recover automatically all the prosodic boundaries detected manually.

Hence, an attempt was made to recover the erroneously detected prosodic boundaries with the two remaining prosodic parameters, that is, with vowel energy and F0 slope.

3.3. Vowel energy use

The vowel energy value was normalized by the mean vowel energy of the recoding and also by the mean vowel energy calculated on the whole training data base for 4 vowel groups: high vowels, low vowels, nasal vowels and mid vowels. Using mean energy values of the four vowel classes aimed to prevent an energy variation due to the intrinsic property of the vowel (low vowels contain higher energy than high vowels). Energy values were modelled using a discrete modelling procedure in three relevant vowel energy positions: ***final position of prosodic groups***, ***initial position of prosodic groups*** and ***other positions***. As for duration, the position of the current vowel was predicted by the model (normalized histograms) corresponding to the greatest likelihood for the observed energy value. The detection results were as follows: about 38% of detected boundaries coincided with boundaries detected by vowel duration and F0 slopes, while 42% of boundaries were erroneously detected and 20% of boundaries were detected only by energy values.

3.4. F0 slope use

F0 slopes were calculated for the vowels in prosodic groups in three positions: ***final position of the prosodic group, final word position*** and ***other positions***. The F0 slope was modelled using a discrete modelling procedure. The position of the current vowel was predicted by normalized histograms corresponding to the greatest likelihood for the observed F0 value. The boundary detection results were the followings: about 40% of detected boundaries coincided with boundaries detected by vowel duration and vowel energy, 55% of boundaries was erroneously detected and 5% of boundaries were detected only by F0 slope values.

4. Discussion

It appeared from the tests that the most efficient parameter to approximate automatically the auditory impression of prosodic boundary detection in French is the vowel duration. However, the vowel duration alone does not lead to very satisfactory results as it does not account for all the auditory features used by listeners especially when eurhythmic patterns are concerned. The two other parameters, vowel energy and F0 slopes, acted as efficient complementary parameters (see Table II). In fact, they allowed to increase the automatic detection of prosodic boundaries and to account for physical parameters that can be used to retrieve automatically prosodic groups in French.

Table II: *Prosodic boundary recovery with energy and F0 slope*

Duration	**Energy**	**F0 slope**	**Total**
67%	***20 %***	***5 %***	***92 %***

5. Conclusions

An auditory analysis of the speech data base showed that consecutive eurhythmic prosodic groups are used by French speakers when speaking spontaneously, creating thus rhythmic patterns. The rhythmic patterns are considered as the repetition of the same length of the prosodic group expressed in number of syllables. The use of the rhythmic patterns turned out to be speaker dependent: some speakers were very keen on using eurhythmicity in their speech production while other speakers used it only in a sporadic way or hardly ever. In this study, the rhythmic patterns of the speech are retrieved by automatic approaches based on prosodic parameter modelling. The results of the prosodic boundary detection are very encouraging and highlighted a hierarchy among the prosodic parameters: the most important parameter in automatic prosodic boundary retrieving was the vowel duration, followed by vowel energy and F0 slope.

6. References

[1] Bartkova, K.; 2005. Prosodic cues of spontaneous speech in French, in *DISS'05*, Aix-en-Provence, France, pp. 21-25.

[2] Bartkova, K.; Segal, N., 2006. Détection automatique de frontières prosodiques dans la parole spontanée, in *JEP 2007*, Dinard, France

[3] Cummins, F.; Port, R.F., 1998. Rhythmic constraints on stress timing in English, in *J. Phonetics*, 26, pp. 145-171

[4] Di Cristo, A., 2003. *De la métrique et du rythme de la parole ordinaire*, Bordas 2003, pp. 25-45

[5] Fónagy, I., 1980. L'accent français, accent probabilitaire, in *Studia Phonetica*, 15, pp. 123-133

[6] Fraisse, P., 1974. *Psychologie du rythme*, Paris : Presses Universitaires de France

[7] Guaïtella, I. 1997. Parole spontanée et lecture oralisée : activités cognitives différentes, organisations rythmiques différentes, in *Travaux de l'Institut de Phonétique d'Aix*, 17 pp. 9-30

[8] Jun, S.A.; Fougeron, C., 2000. A phonological model of French intonation, in *Botinis, A. Intonation : Models and Technology*. Dordrecht : Kluwer Academic Publishers, pp. 209-242

[9] Lehiste, I., 1977. Isochrony reconsidered, in *Journal of Phonetics*, 5, pp. 253-263

[10] Pasdeloup, V. 2004. Le rythme n'est pas élastique : étude préliminaire de l'influence du débit de parole sur la structuration temporelle, in *JEP*, Fès, (Maroc), 2004, pp. 397-400

[11] Pike, K. 1946. *Intonation of American English*, Ann Arbor, *MI* : University of Michigan Press

[12] Pompino-Marschall, B., 1989. On the psychoacoustic nature of the P-center phenomenon, in *Journal of Phonetics*, 17, pp.175-192

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