Automatic phrasing in French

Philippe Martin¹

¹LLF, UFRL, Université Paris Diderot
philippe.martin@linguist.univ-paris-diderot.fr

Abstract

Whether we read aloud or silently a text in lexically stressed languages such as Italian or English, we segment speech not in words, but in sequences containing a content word (noun, adverb, verb or adjective) together with its associated grammatical word(s). These sequences are called accent phrases, and contain a single pitch accent whose location on the content word is defined in the lexicon.

The simple fact that we can restore lexical stress in silent reading suggests that we don’t really need the actual presence of specific acoustic features of speech, such as vowel duration, fundamental frequency change or intensity modulation, often mentioned in the literature as parameters of stress.

In French, the actual phrasing, i.e. the segmentation into accent phrases, depends strongly on the speech rate adopted by the speaker or the reader. Using a slow speech rate, all words could be stressed on their last syllable, whereas a fast speech rate could merge up to 10 syllables together in a single accent phrase containing more than one content word.

Based on various characteristics of accent phrases in French, an algorithm operating in a top-down fashion for automatic identification of stressed syllables is described and applied on examples of read and spontaneous speech.

Index Terms: Accent phrase, phrasing, French, stress location, automatic segmentation.

1. Introduction

When we read a text, either aloud or silently, we could proceed word by word, or even syllable by syllable, but if we master the language and identify all the words, we usually proceed by group of words. In the following example, each of these groups carry a single stressed syllable placed on some syllable of a content word (a noun, a verb, an adverb or an adjective) as defined in the lexicon: [in the orthographic] [representation] [of speech] [of most] [written] [languages] [segmentation] [is defined] [by spaces] [between] [words]. Such groups of words are called in prosodic phonology accent phrases, and define the minimal prosodic units, which organized into a hierarchy, constitute the prosodic structure of the sentence [1], [2].

For all fluent speakers of English, the placement of these stressed syllables on content words in accent phrases is predictable, and results from the acquisition of the lexicon. Other stressed syllables can also occur in speaker’s production, but contrary to lexically defined stress, they are not predictable as they result from a specific choice of the speaker to indicate an emphasis. Emphatic stress may occur on a different syllable than lexical stress or on the same syllable. In this latter case the speaker will use a different acoustic realization as emphatic stress has to be perceived by listeners as different and phonologically unpredictable than the predictable lexical stress.

Phrasing determines an essential step in the comprehension of speech. The segmentation into accent phrases constitute the first step to rebuild the prosodic structure intended by the speaker or the writer and access at a later stage the syntactic structure. In order to read a text silently, we do generate mentally a prosodic structure, as indicated by (mental) melodic contours located on stressed vowels, although any acoustic input is of course absent.

The simple fact that we can restore stress locations when we read aloud or silently suggests that we may not really need any acoustical input to perceive stressed syllables when we listen to speech (again non-emphatic). Not only reading aloud or silently of the same text could lead to different phrasings, but the predictability of lexical stress may lead to consider the perception of stressed syllables as the result of a process which would compare the actual acoustic features of syllables with a position predicted from the knowledge of the lexicon. When we listen to speech, in a process similar to the one used in silent reading, we could eventually restore stress mentally in a location where we would have placed the stressed syllable ourselves (cf. recent studies [10], [18], [19]). Conversely, the ignorance of the language would prevent the identification of stressed syllables in a language even if the relevant acoustic parameters are present, as demonstrated in the perception experiment conducted on Berber and Hebrew [3].

2. The case of French

In French the position of lexical stress evolved gradually to the last syllable of content words by progressively dropping all post stressed syllables [4]. The function of lexical stress as marker of morphological boundary as in the other lexically stressed Romance languages was gradually lost as redundant. It became then possible to skip some of the expected stress locations when speaking, as in la petite armoire violette “the little purple cupboard”, which can receive one, two or even three stressed syllables: la petite armoire violette, la petite armoire violet. la petite armoire violette or la petite armoire violet. For a French speaker, it is easy to notice that the difference in stress patterns and therefore in phrasing in this example is linked to speech rate. In order to pronounce (or even to read silently) la petite armoire violette with only one final stressed syllable on violette, one has to use a (very) fast speech rate, whereas a slower pace would lead to the pronunciation of three stressed syllables as in la petite armoire violet.

We could perhaps then conclude that there is no limit to the number of syllables and words that can be pronounced in French with only one final stressed syllables, i.e. in a single accent phrase. The pronunciation of long words will help discover where the limit stands. Long words such as the well-known anticonstitutionnellement “against the constitution”, (8
syllables) or intergouvernementalisation “inter
governmentalization” (10 syllables) seem difficult if not
impossible to pronounce or even silently with only one
final stressed syllable. Already in 1550, Louis Meigret [5]
proposed that the longest word that could be pronounced with
only one final stress would have a maximum of 7 syllables.
Later, Martin [6] showed that it was not the number of syllables
that matters, but the time needed to pronounce them. The data
obtained from fast speech rate speakers show that the maximum
interval between consecutive stressed syllables (in flowing
speech) could not exceed some 1250 ms (depending on the
subjects). In parler jeune productions in French (the young
people speaking style), sequences of up to 10 or 11 syllables
with only a final stress have been observed (Lehka and Le Gac
[7]).

While 1250 ms (the value retained here for the sake of
discussion) is the maximal duration between consecutive
stressed syllables in connected speech, a minimal duration
between two consecutive stressed syllables can also be
established and will indirectly define a minimal duration of
accent phrases therefore containing then only one syllable,
necessarily stressed, and necessarily preceded by a sufficient
silent gap. If not, the preceding syllable will be perceived as
unstressed whatever its acoustic characteristics and will belong
to the AP ended by the stressed syllable. This limit is easy to
establish experimentally, by selecting natural or synthetic cases
of consecutive stressed syllables, as for example par le fait que
or le travail de nuit nuit i.e. cases of stress clash with no move
or deletion of the first stress. Indeed, it is often mentioned in
the literature that these cases require a kind of acoustic gap between
consecutive stressed syllables (Di Cristo, [8]), usually but not always
implemented by the presence of some consonants after
the first of before the second stressed syllable. However, it is
experimentally easy to reduce the gap with a sound editor until
the first implied syllable ceased to be perceived as stressed
while nothing of its acoustical structure has been modified (i.e.
by removing the silent part only). This limit, measured vowel
perception center to the next vowel perception center, is about
250 ms, which implies that the minimal duration of an accent
phrase must include some silence added to the duration of its
unique vowel, as below this value, the word owning the first
syllable will become part of the newly formed accent phrase.
For example, the perceived desaccentuation of fait in [par le
fait] [que] will merge the accent phrase par le fait with the
second accent phrase que to form the new group [par le fait
que]. The minimal duration between consecutive stressed
syllables is thus about 250 ms.

3. Syllables followed by silence are stressed

It is equally easy to demonstrate experimentally that any
syllable followed by at least 250 ms of silence is perceived as
stressed. Either by inserting some 250 ms or more acoustic
silence on the speech wave using a sound editor without
modifying the acoustic characteristics of the syllable, or by
slowing the speech rate so that the number of syllables reaches
a level below some four syllables per second, the final syllables
of any word become perceived as stressed, whatever their actual
duration or pitch movement [9].

The actual explanation for this perception effect is linked to
the processing of syllables by the brain, and more precisely by
the brain oscillations carrying information between neuronal
zones (Martin, [10]). It can be shown that the perception of
syllables needs at least 100 ms processing time, even if their
actual duration is below this value (Ghitza, [11]). If given more
than some 250 ms, a normally unstressed syllable becomes
perceived as stressed, without changing its acoustic structure.
Since two consecutive stressed syllables must be separated by
at least 250 ms, we can conclude that the perception of stressed
syllables needs at least 250 ms processing time. Incidentally,
this property is used in an automatic detector of (some) prominence in speech in Analog [12].

4. Eurhythmy

The eurhythmocity observed for both read and spontaneous speech may also be taken into account in a top-down approach for prosodic segmentation. Spontaneous speech eurhythmity proceeds frequently by adjusting the average duration of accent phrases syllables to reach comparable duration of successive accent phrases. Read speech uses more often a strategy aiming to balance the number of syllables of successive accent phrases, at the possible expense of congruence with the syntactic structure (Wioland, [13]). A classic example is given by a sentence such as Marie adore les chocolats “Mary loves chocolates” where spontaneous speech subjects would have a tendency to realize a phrasing congruent with syntax [Marie
[adore les chocolats]] and possibly aim for eurhythmity by
slowing the syllabic rate of [Marie] and going faster on [adore
les chocolats]. On the contrary, readers of this sentence show a
tendency to group the words to balance the number of syllables in consecutive accent phrases, at the expense of congruence with syntax [Marie[adore] les chocolats].

5. Glissando

An interesting characteristic of pitch accents pertains to their instantiation as melodic contours. Although in the Autosegmental-Metrical approach pitch accents are not considered as interacting with each other, other models describe them as indicators of dependency relations existing between accent phrases [9]. Using a terminology borrowed from Delattre [14] but applied to stressed vowels only and not to the global pitch curve, minor continuation, instantiated by a falling contour, indicates a dependency towards a major continuation instantiated by a rising contour, and a major continuation indicates a continuation towards the terminal conclusive contour. These relations of dependency “to the right” (i.e. toward an accent phrase that would follow in the sentence) are indicated by a contrast of melodic slope, where the falling pitch minor continuation indicates a dependency toward a rising pitch major continuation.

Obviously, this model implies that the falling and rising melodic slope are effectively perceived, i.e. that the speed of melodic change in time is above what is called the glissando threshold and that the melodic movement is perceived as a change in frequency and not as a static tone. The glissando threshold is evaluated as the difference from the beginning to the end in semitones referred to the duration of the contour, assuming a linear variation, cf. Rossi, [15].

6. Stress predictions

While listening to speech, we cannot prevent to have expectations towards the location of stressed syllables different
from the one actually realized by the speaker. In French, these expectations depend on the adaptation to the speaker speech rate.
This adaptation is not always easy or even possible. The examples provided by the parler jeune with a very fast speech rate exceeding 8 or 9 syllables per second are hard to match for most listeners, to the point that some native speakers will have trouble to understand such speech tempo. Therefore, some listeners will have a tendency to hear stressed syllable even if they do not exist acoustically. In the following example, the actual accent phrases acoustically realized by the speaker are marked in red while potentially stressed syllables are marked in blue:

[C’est toi qui a pris la responsabilité de casser] pronounced with 14 syllables

[Je vois pas pourquoi tu viens casser les couilles aux gens] 12 syllables

The first accent phrase contains 14 syllables which is unusual for the average speaker of French. Therefore, any listener not practicing the parler jeune will have a strong tendency to restore mentally a stressed syllable on toi leading to a different phrasing ([C’est toi] 2 syllables [qui a pris la responsabilité de casser] 12 syllables (elision of [a] in qui a), or possibly on pris et responsabilité, resulting in a four accent phrases phrasing ([C’est toi] 2 syllables [qui a pris] 2 syllables (pronounced [kapri]), [la responsabilité] 7 syllables, [de casser] 3 syllables. Still, the only obvious acoustical marker of stress is on the final syllable of responsabilité.

Likewise, the second accent phrase of the example with 12 syllables, could be mentally segmented into 2 or 3 accent phrases, depending on the speech rate adopted mentally or in oral production: [Je vois pas pourquoi] [tu viens casser les couilles] aux gens (the [s] of the pronoun je is deleted here).

7. Stress detection: Mission impossible?

The problem for an annotator of stressed syllables in French is to adapt to the speech rate of the recording when accentuated syllables are annotated. The perception of stress will be influenced by the annotator own prediction process, tending to detect accentuated syllables there would have placed by reading or speaking at the annotator own pace.

Traditionally, automatic detection of stressed syllables in French operates in a bottom-up fashion from the speech acoustic data, looking for significant variations between consecutive syllables in duration, fundamental frequency and intensity. Vowel quality does not appear as a significant parameter for stress detection in French.

M. Avanzi [16], faced with the uncertainty in annotating stressed syllables in French, describes in detail a complex procedure involving two experts, possibly helped by a third in case of disagreement between the two first experts. Even with this protocol, agreement between annotators varies between 60 % and 80 %.

In another paper on the same subject, Christodoulides and Avanzi [17] implemented an automatic detector of prominence (i.e. not just accent phrases stressed syllable) by machine learning methods applied to a large corpus which included two different styles. Although they use a comprehensive set of acoustic parameters, their best results, evaluated against manual placement by experts in syllabic prominence, reaches a 90% correct identification level.

Considering these difficulties, and the observations mentioned above, it appears that stress detection for phrasing should proceed not directly from speech wave analysis, but rather from the knowledge a system could have access to beforehand, the location of potential positions as final syllables of content words among other.

To attain a reasonable chance of success, a computer implementation dealing with speech wave should adopt a comparable strategy, and not infer results starting from acoustical analysis of speech but rather from the expected stressed syllable locations. The availability of transcribed and segmented speech down to the syllabic and phone level data should be a prerequisite towards automatic stress detection, as the syllabic candidates for stress can be directly inferred from the aligned transcription.

8. Computer implementation

To apply to French the definition of accent phrases given for lexically stressed languages, we assign a stressable label to final syllables of all nouns, verbs, adverbs and adjectives. To select actual stressed syllable among the list of stressable ones, we use the constrains described above, i.e. the minimal and maximal duration of accent phrases (respectively 250 ms and 1250 ms), the minimal separation of 250 ms between two consecutive stressed syllables, and the presence of at least 250 ms of silence following a stressed syllable. The application of these constrains would make some stressable syllables actually stressed (in a time window of 1250 ms for example) and eliminate some from the list of possibly stressed.

The next step for the selection of actually stressed syllables, without even looking at the speech wave, would be to use the speech rate, i.e. the number of syllables per second evaluated from the corpus text transcription. Linked to an average number of syllables per accent phrase, we can have an approximation of the phrasing realized in a given recording, validated by an assumed eurhythmicity.

According to this definition of accent phrases as minimal units of prosody whose hierarchy constitute the sentence prosodic structure, we can assign to any stressable syllable whose change in fundamental frequency on its vowel exceeds the glissando threshold as actually stressed. Although this step assumes the linearity of the fundamental frequency change of the syllable vowel as well as the accepted validity of the glissando threshold (which in fact is parametric), we have enough tools to implement an innovative algorithm for automatic selection of stressed syllables from a list of stressable syllables.

9. Automatic detection of stressed syllables in French

From these various observations and considerations, the following rules for a computer implementation can be applied:

1. Any syllable followed by more than 250 ms silence is (perceived as) stressed
2. Any final syllable of a noun, adjective, verb or adverb is stressable (from accent phrase definition)
3. If 2 consecutive stressed syllables are separated by less than 250 ms, the first one is unstressed (accent phrase minimum duration)
4. Any stressable syllable with F0 change over the glissando threshold is stressed
5. If 2 consecutive stressed syllables are separated by more than 1250 ms in continuous speech, at least one stressable syllable in this interval is stressed (accent phrase maximum duration). Make stressed the one with the highest glissando value

6. One stressable syllable must exist in any time window duration equals to the accent phrase average duration (eurythmy)

The eurhythmic aspect is implemented by evaluating the first accent phrases realizations and the number of syllables they contain. This starting accent phrase duration will then be used to define a sliding time window, in which most prominent syllables in value of glissando will be retained as stressed. The size of this sliding window defines a speech rate assumed to be constant in the whole recording.

The following read example illustrates this point.

Il était une fois un pauvre escargot qui souffrait beaucoup à chaque fois qu’il partait en randonnée

Step 1: Any syllable followed by more than 250 ms silence is stressed (in red)

Il était une fois un pauvre escargot qui souffrait beaucoup à chaque fois qu’il partait en randonnée

Step 2: Any final syllable of a noun, adjective, verb or adverb is stressable (in blue)

Il était une fois un pauvre escargot qui souffrait beaucoup à chaque fois qu’il partait en randonnée

Step 3: If 2 consecutive stressed syllables are separated by less than 250 ms, the first one is unstressed

Il était une fois un pauvre escargot qui souffrait beaucoup à chaque [180 ms] fois qu’il partait en randonnée

Step 4: Any stressable syllable with F0 change over the glissando threshold is stressed (value/threshold with coeff. 0.16)


Step 5: Two consecutive stressed syllables separated by more than 1250 ms, this is the case for the last accent phrase

à chaque fois qu’il partait en randonnée 1367 ms

We can select the highest glissando value

à chaque fois qu’il partait en randonnée 1367 ms or both stressable syllables

à chaque fois qu’il partait en randonnée 1367 ms

Step 6: Apply eurhythmicity to retain the latter possibility

726 ms 5 syl. 145 ms/syl. Il était une fois
687 ms 5 syl. 137 ms/syl. un pauvre escargot
765 ms 5 syl. 153 ms/syl. qui souffrait beaucoup
407 ms 3 syl. 135 ms/syl. à chaque fois
487 ms 3 syl. 162 ms/syl. qu’il partait
546 ms 4 syl. 136 ms/syl.en randonnée

The average accent phrase duration is about 709 ms.

Another example this time of spontaneous parler jeune:

Step 1: Juste pour une carte d’identité t’as pas ta carte tu fais tes vingt-quatre heures tu ressors t’as la haine encore plus ça augmente

Step 2: Juste pour une carte d’identité t’as pas ta carte tu fais tes vingt-quatre heures tu ressors t’as la haine encore plus ça augmente

Step 3: Juste pour une carte d’identité t’as pas ta carte tu fais tes vingt-quatre [230 ms] heures tu ressors t’as la haine encore [240 ms] plus ça augmente


Steps 5 and 6 do not apply

227 ms 1 syl. 227 ms/syl Juste
356 ms 3 syl. 118 ms/syl. pour une carte
537 ms 4 syl. 134 mr/syl. d’identité
569 ms 4 syl. 142 ms/syl. t’as pas ta carte
945 ms 6 syl. 157 ms/syl. tu fais tes vingt-quatre heures
486 ms 3 syl. 162 ms/syl. tu ressors
431 ms 3 syl. 143 ms/syl. t’as la haine
496 ms 3 syl. 165 ms/syl. encore plus
592 ms 3 syl. 197 ms/syl. ça augmente

The average accent phrase duration is 515 ms.

Some cases may pose problem, such as examples with a tonic pronoun placed after the verb. If the stress pattern of Redonne [moi la main] is correctly predicted by the algorithm, but leads to an unexpected accent phrase [moi la main], Redonne moi la main being emphatic, the stress pattern Redonne [la main] [plus loin] is quite possible and leads to consider some pronouns as stressable even if they are not followed by 250 ms of silence.

10. Conclusions

The following and accent phrase properties are considered to implement a top-down automatic phrasing segmentation in French: 1) Speakers and readers of French are capable to restore accent phrase stressed syllables even without any acoustic input 2) The minimum duration of accent phrases is 250 ms, and the maximum about 1250 ms 3) The actual duration of accent phrases depends on the speech rate selected by the speaker of the reader 4) The actual syllabic stress defining phrasing carries a melodic movement above the glissando threshold.

11. References


