How far can prosodic cues help in word segmentation?

Katarina Bartkova

France Telecom R&D
Lannion, France
Katarina.Bartkova@rd.francetelecom.com

Abstract

Prosodic cues are of great importance in parsing speech signal into prosodic and lexical units. Listeners detect the changes of the prosodic parameters and interpret them to detect sentence modalities or the mood of the speaker. Some automatic speech recognition systems try to use prosodic parameters to detect boundaries of prosodic units and help thus the acoustic decoding process. Although the automatic detection of major prosodic boundaries is most of the time reliable, minor boundary detections are prone to error. However, listeners, unlike automatic processing systems, can detect with great precision boundaries of lexical items even if they do not coincide with major prosodic boundaries. Our feeling is that a deeper understanding of the prosodic parameters in spontaneous speech would improve their modeling and ultimately their use by automatic systems. This study analyses filled and silent pause occurrences and two prosodic parameters, duration of pauses and vowels and F0 slopes, measured on a spontaneous speech corpus in French. The results of the analysis revealed that a simple local comparison of the parameter values with the values measured in the vicinity of the segment under consideration can provide valuable information on the lexical boundaries as well as on prosodic patterns of the lexical units.

1. Introduction

Prosodic features reflect the syntactic or semantic cohesiveness of the successive words. Juncture between syntactic constituents is made by prosodic means such as appropriate adjustment of the pitch contour and phoneme durations or pause insertions. Listeners make use of these cues to segment the incoming flow of speech into words.

One of the major problems encountered when modeling prosodic parameters is that the prosodic phrase structure is not deterministic, that is speakers can produce a sentence in several ways without altering the naturalness of the meaning. The mapping between syntax and prosody is complex because according to [3], in speech, the constraints of syntactic structure and phrase length are balanced to produce a regular sequence of prosodic phrases.

The use of prosodic parameter is crucial in automatic speech processing techniques. Lately, modeling prosodic parameters became of great interest also in automatic speech recognition as this technique focuses ever more on the recognition of fluent and spontaneous speech. While phone duration or energy were in isolated word recognition tasks mainly used as prosodic cues (e.g. for detecting an inappropriate duration of the speech signal) [2], in fluent speech recognition, prosodic parameters can be fully employed to detect the demarcation of the prosodic units. The detected prosodic boundaries can then be used during the speech recognition to constrain the search space [5]. However, boundary detections are generally limited to major boundaries often accompanied by pauses. This fact is mainly due to the rather low detection rates, as minor boundaries are not accompanied by pauses.

This study attempts to shed light on the characteristics of the demarcative information encoded by prosodic means. Silent and filled pause occurrences are analyzed, the congruence between silent pause occurrences and syntactic juncture is questioned and the prosodic cues of filled pauses are studied. The second part of the paper addresses the analysis of the prosodic cues when silent pauses do not follow lexical units.

2. Method Overview

2.1. Speech data base used

The speech data base used here is constituted of more than 1080 telephone messages in French, left by clients as the result of a survey dedicated to the analysis of client's satisfaction. The data base contains 55180 words, which means that on average 54 words are uttered per message. The data base was manually transcribed including the annotation of non-speech noises such as respiration, laughter, background noises… as well as interrupted words, interrupted sentences and filled pauses. This orthographic transcription was used to automatically align the words and their phonetic transcriptions with the speech signal. This way it became possible to access prosodic events such as occurrences of silent and filled pauses durations, the F0 values and vowel durations.

2.2. Investigated prosodic parameters

In order to characterize the speech at the prosodic level, prosodic values such as phone duration and F0 slopes were measured on every vowel of the lexical units, except final schwa-like vowel, which can occur in French after each uttered consonant even when not present in the spelling form. The parameters associated to the schwa vowel were used exclusively when it was the only vowel in the word.

The F0 slope was calculated for each vowel under consideration as the difference between the F0 measured at the end and at the beginning of the vowel (no normalization according to the vowel length was carried out). In order to make the slope comparisons easier, five categories of F0 movements were differentiated in this study: a flat movement when the slope of the F0 value did not exceed -3 and +3 Hz, an upward movement when the slope's value was above +3 Hz and below 15 Hz and a steep upward movement when the slope value was higher than 15 Hz (henceforth termed as mid-high and high-high slopes). The same distinction was made for negative slope values: an F0 slope which was between -3 and -15 Hz, was considered as a moderate downward slope (termed as mid-low) and when it was lower than -15Hz, it was

categorized as a deep low slope (henceforth termed as low-low).

To facilitate the statistical analysis of vowel durations, they were grouped into 50 ms intervals.

Silent pauses were categorized according to their durations into three categories: short pauses corresponding to pauses shorter or equal to 150 ms, long pauses when longer than 300 ms and mid-long pauses in between (i.e. from 150 to 300 ms).

A distinction was made between function and lexical words as prosodic features can affect differently these word categories [6]. Each prosodic parameter value measured on a vowel was compared with the values of the same parameter measured on vicinity of the vowel under consideration.

It is often claimed that the major prosodic cues of syntactic structuring in French are located mainly on the last syllable of the word. It has been detected that [4] that low and high F0 slopes alternate on syntactic junctures in reading style. As far as vowel duration is concerned, they are lengthened on syntactic borders as a function of the boundary depth and of the right consonantal context [1]. However, the length of the last vowel, located on syntactic juncture, is relatively independent from duration of the other phones of the same prosodic group.

3. Prosodic parameters associated to pauses

Pauses are vital in speech production for the speaker, as he must breathe, but they are also necessary for listeners, since they provide time to decode the incoming flow of speech. In fluent speech, pauses are situated on syntactic boundaries to indicate the syntactic parsing of the speech or are used for stylistic purposes to enhance the word meaning. However, as it appears from our data, pause distribution in spontaneous speech does not always reflect congruence between syntax and prosody.

3.1. Silent pauses

In our data 8346 speech internal silent pauses were detected. On average a short or an intermediate pause followed every 4th word and a long pause every 5th word.

As explained before, pauses are grouped into 3 categories: short, intermediate and long. The distribution of the pauses was the following: **45%** (3728) were **short pauses** (shorter than 150 ms), **14%** (1137) were pauses of **intermediate length** (from 150 to 300 ms) and **42%** (3481) were **long pauses**, i.e. longer than 300 ms.

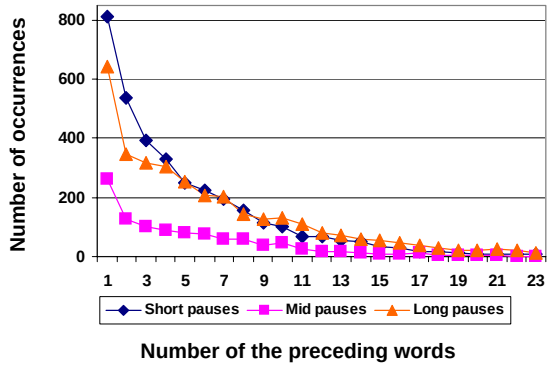

Figure 1: Number of pauses as a function the number of preceding words

Figure 1 illustrates the number of pauses as a function of the number of words preceding the pause. It appears for all three categories of pauses that their occurrence after one single word accounts for the highest amount of occurrences. Such a distribution of silent pauses suggests that speakers can use them as a talk preparation gap and that pauses are not necessarily situated on syntactic boundaries. This is partly confirmed by the F0 patterns measured before pauses and reported in Figure 2. As it was previously recalled, in French, the F0 pattern on prosodic boundaries has a clear falling or rising movement. However, a large amount of F0 movements, measured on the last vowel preceding the pauses, were flat. Therefore, it can be assumed that pauses preceded by a small number of words and by a last vowel with flat F0, occur where hesitation is present.

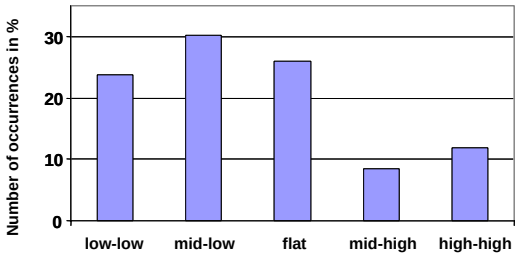

Figure 2: F0 pattern preceding a silent pause

3.2. Filled pauses

As the corpus used in this study is constituted of spontaneous speech, disfluencies such as filled pauses frequently occurred in the speech signal. Filled pauses should be filtered out from the speech before the semantic interpretation of the recognized items. Filled pauses, though relatively well detected on the acoustic level, are still often substituted with other vocabulary words. Prosodic characteristics of the filled pauses are very robust; therefore, they could be used with high confidence by automatic recognition systems.

There were 2871 filled pauses in our corpus. The filled pauses could be separated from the surrounding words by a silent pause (in 32% of cases). They could be attached to the preceding word (in 68%) as a very long schwa-like vowel uttered after a final consonant. They could also be perceived as a neutralized part of the preceding vowel timber when attached to the last vowel of a previous word.

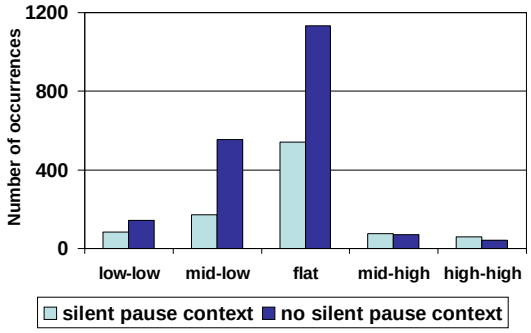

Figure 3: F0 pattern in filled pauses

The main prosodic cue of a filled pause is its long or very long duration (the mean duration measured on our data was equal to 350 ms) and its F0 movement which was mainly flat. Figure 3 illustrates the prosodic cues of filled pauses

surrounded or not by silent pauses. Although the F0 patterns were mainly flat, the downward F0 pattern, especially a moderate one (mid-low), was also quite frequent.

4. Prosodic parameters of lexical units

The speech signal exhibit prosodic parameters that enable listeners to break the signal down into lexical units. The segmentation performance of automatic devices is far from listener's one since problems of word segmentation occur whenever no clear boundary cues are present in the speech signal. If prosodic parameters are to be used by automatic systems to help the segmentation of lexical units, they ought to be robust enough in order produce clear distinctions. The demarcation of lexical or syntactic units is carried out by prosodic means, mainly with an increase in F0 value, syllable duration and phone energy. In French, where strictly speaking lexical stress does not exist, the prosodic demarcation coincides with the edges of the lexical units. Major prosodic boundaries exhibit large changes of prosodic parameter values and generally they are relatively easy to localize. But how reliable can be the detection of minor boundaries? In order to answer this question, vowel duration and vowel F0 slopes were analyzed in words not followed by pauses and the results are presented in the following section.

4.1. F0 pattern

Similarly to vowel durations, a local comparison of F0 slope values was carried out between the slope of the vowel under consideration and the slopes of the previous and following vowels. When the direction of the F0 slope is the same on the last vowel of a word as it is on the first vowel of the next word, then the slope is considered as embedded in a larger F0 movement. In such a case the F0 parameter alone cannot provide the reliable prosodic cue for lexical boundary detection. On the other hand, when the slope movement measured on the last vowel is different form the one measured on the first vowel of the following word, then the slope can be considered as a good candidate to indicate the lexical frontier.

The following analysis tries to shed light on the prosodic articulation between the current and the following word and on the possibility to detect, when possible, typical word slope patterns. As inner-word F0 patterns comparison is impossible in mono-syllabic words, a separate analysis is applied to mono-syllabic and pluri-syllabic words.

Mono-syllabic words

More than 50% of the words of the corpus were mono-syllabic words. A distinction was made here between lexical and function words as the word grammatical category can influence its F0 pattern.

Around 40% of the mono-syllabic words were function words. Function words had seldom rising F0 slopes (see Figure 4). They have most of the time flat F0 slopes (26% of the cases), moderate negative slopes (37 % of mid-low) and steep negative slopes (21% of low-low). The F0 slope measured on the vowel in function words is most of the time embedded in a larger F0 movement imposed probably by the word to which the function word is associated. However, changes in F0 movement direction can still intervene after a function word (see Figure 5). Such a change is generally associated to a flat or a moderate downward (mid-low) F0 movement.

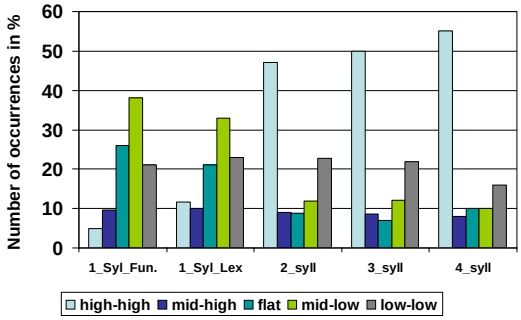

Figure 4: F0 slope measured on the last vowel of word as a function of the word length

One syllable content words have slightly more high F0 slopes than function words (11 % compared to 5 % in function word, see Figure 4). Apart from this distinction, the F0 slope values of lexical and function words were quite similar.

Pluri-syllabic words

25% of the words of the corpus had 2 syllables, 10% of the words had 3 syllables and 3% 4 syllables. Words longer than 4 syllables were not analysed as their low number did not allow us to obtain statistically reliable results.

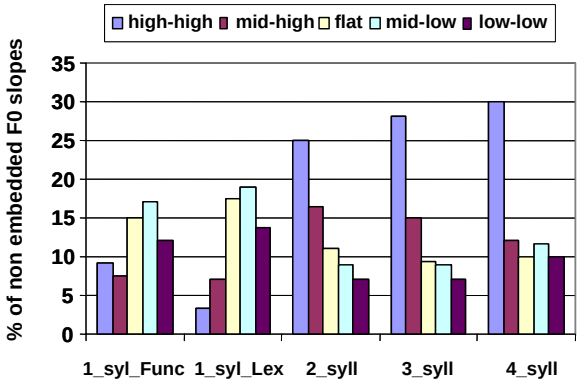

Figure 5: Proportion of non-embedded slopes for each type of F0 slope and word length

The values of F0 slopes measured on the last vowel of pluri-syllabic words were quite different from those obtained on mono-syllabic words. In fact, as it appears in Figure 4, longer words had a high number of upward steep (high-high) F0 slope. The number of their downward extreme slopes (low-low) is the second most important one (although clearly lower in number than the number of the extreme high slopes).

Moreover, the final F0 pattern of pluri-syllabic words is less often embedded in a general F0 movement than the F0 pattern of mono-syllabic words (see Figure 5). A change in slope direction occurs frequently after a rising F0 movement, especially after a high-high F0 one.

The relation of the F0 slopes measured on penultimate and on last vowels of the words was also analyzed. In doing so we tried to investigate the amount of F0 slopes of the penultimate vowels embedded in a general pattern of the word final F0 movement. Figure 6 indicates the F0 slopes on both the penultimate and last vowels of pluri-syllabic words. The figures indicate, for each penultimate vowel, the percentage of by upward and downward F0 movements measured on the last vowel of the word. Since the results obtained for the different lengths of pluri-syllabic words were relatively

similar, the cumulated results are represented in a single diagram.

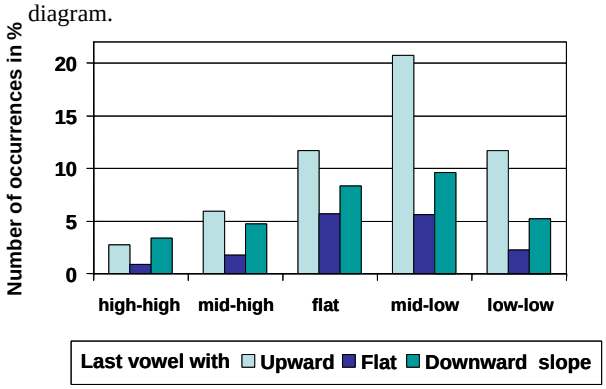

Figure 6: F0 slopes on penultimate vowels followed by a last vowel rising, flat or falling F0 slope.

As it can be seen in Figure 6, rising F0 slopes measured on the penultimate vowel are often embedded into a general upward movement typical for final word position. However, when the penultimate F0 pattern is a falling one, most of the time a direction change is carried out bringing about a rising F0 pattern on the last vowel.

As far as an overall F0 word pattern is concerned, it appeared from the data analyzed that the longer the word is and more it favors rising final F0 slopes. There were very few examples, if any, in our data of long words (4 syllable ones) with flat slopes on penultimate and last vowels preceded by rising or falling F0 patterns.

4.2. Vowel duration

The duration of the last syllable, especially when it is stressed, is longer than the same syllable in an unstressed position. When the duration of a syllable is longer than the duration of the following one, a lexical boundary can be suspected. The need for longer syllable duration indicating the word boundary is increased when the movement of the F0 pattern of the last syllable of a word is embedded in a general same directional F0 movement.

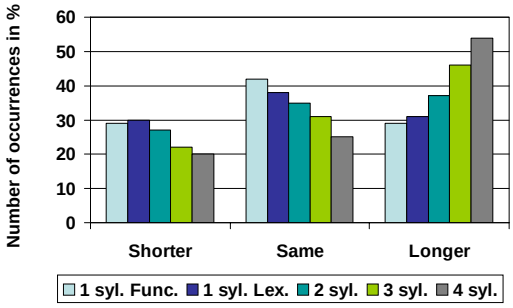

Figure 6: Last vowel duration as a function of word length compared to the first vowel duration of the following word

Figure 6 reports the duration of last vowels of words whose F0 pattern is embedded with the pattern of the next vowel. As it appears from the data, there is a correlation between vowel length and word length (expressed in number of syllables): the longer the word, the higher the probability that its last vowel duration will be longer than the first vowel duration of the following word.

As far as the last vowel duration comparison with the penultimate vowel duration of the same word is concerned, it appeared from our data, that the last vowel is very seldom shorter than the penultimate one; most of the time, the last vowel has a similar or longer duration than the penultimate vowel (see Figure 7). Again a slight correlation with respect to the word length (in syllables) exists: the last vowel of a long word (in terms of syllables) is more often longer (compared to the penultimate vowel) than the last vowel of a short word.

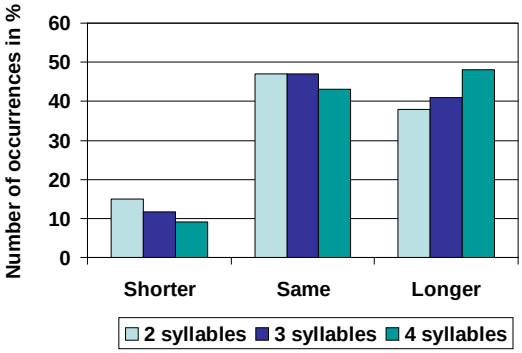

Figure 7: Comparison of the duration of the last and penultimate vowel of word as a function of the word length

5. Conclusions

The present study addressed the problem of prosodic parameters in words not followed by pauses. It aimed to investigate how typical prosodic parameters are in order to be used successfully by automatic recognition systems. Our results indicates that demarcation cues yielded by vowel durations and F0 slopes could be successfully detected when carried out a simple local comparison of the parameters measured on a few number of successive vowels.

It should be investigated, as a possible continuation of this study, what role is played by the third major prosodic parameter, by phone energy, in the demarcation procedure. Moreover, it also should be investigated how our findings reported here can be successfully integrated into an automatic speech recognition system in order to segment the speech signal into lexical units.

6. References

[1] Bartkova, K., Sorin C.; 1987. A model of segmental duration for speech synthesis in French, *Speech Communication* 6, pp 245-260.

[2] Bartkova, K., **2000.** Prosodic parameters of French in a Speech Recognition System, *Text, Speech and Language Technology : Intonation : Analysis, Modelling and Technology*, A. Botinis (ed.), Kluwer Academic Publishers, pp 357-382;

[3] Gee,J., Grosjean F.; 1983. "Performance structures: A psycholinguistic appraisal. *Cognitive Psychology* 15:411-458.

[4] Martin P.; 1981. Pour une théorie de l'intonation, *L'intonation de l'acoustique à la sémantique*, Klincksieck Paris pp. 234-271.

[5] Niemann H. et al.; 1998. "Using prosodic cues in spoken language systems", *Proc. SPECOMWorkshop, St. Petersburg*, pp.17-28.

[6] Vaissière J.; 1997. Langue, prosodies et syntaxe, *Revue Traitement Automatique des Langues*, Vol.38, 1 pp.53-82.