In search of intonational cues to content word beginnings in conversational speech

Pauline Welby, Robert Espesser, Christine Meunier

In search of intonational cues to content word beginnings in conversational speech

Pauline Welby, Robert Espesser, Christine Meunier

Laboratoire Parole et Langage (CNRS and Université de Provence), France

pauline.welby@lpl-aix.fr, robert.espesser@lpl-aix.fr, christine.meunier@lpl-aix.fr

Abstract

We used an annotated conversational French speech corpus to 1. investigate whether the intonational rises that occur at the beginning of French content words in read speech (APRs) are also present in spontaneous speech and therefore available as cues to word segmentation and lexical access, and 2. test two measures of characterizing intonation patterns using automatically extracted F0 and time values. The two measures tested both proved problematic: they were sensitive to the segmental composition of the critical region. We found no evidence that APRs are reliably present in the corpus as a whole, although we suggest that they may be present in particular types of conversational speech.

Index Terms: intonational cues, word segmentation, lexical access, conversational speech

1. Introduction

There is growing evidence from a number of languages that listeners use intonational patterns as cues to word segmentation, the process of locating word boundaries in the speech stream, and to lexical access, the retrieval of words from the mental lexicon. In the present study, we examine whether the intonational cues shown to be present in earlier studies of French read speech are also present in spontaneous, conversational speech.

French accentual phrases (APs) that are not utterance-final are typically realized with fundamental frequency (F0) rise whose peak is aligned to the last syllable of the AP (the late rise, final rise or primary accent). If an AP is long enough, another rise is often, but not always, realized toward the beginning of the AP. (see [8,17,18] and references therein). This early rise (initial rise or secondary accent) does not typically contribute to the meaning (but see [1]). In short APs, a single rise combining the characteristics of the late rise and the early rise may be realized.

Welby (2003, 2007) [17,19] showed that French listeners use the early rise to locate content-word beginnings in an offline identification task, as suggested by [15,6]. Listeners interpreted sequences like [me.la.mo~.din] as a single 4-syllable content (non)word *mélamondine* when the early rise began at the first syllable [me] and as a proclitic function word followed by a 3-syllable content (non)word *mes lamondines* 'my lamondines' when the rise began at the second syllable [la].

These results were recently extended [13,14], using an online task (cross-modal priming with lexical decision) and resynthesis of the F0 of real word stimuli to show that these AP-initial boundary rises (APRs), which include both early rises and the single rises found in short APs, speed lexical access (see also [12]).

Researchers working on other languages have also found evidence for the use of intonational cues to word boundaries (e.g., [9] for Korean, and [16] for Japanese), and these studies include analyses of speech corpora.

The earlier studies of French are all based on read speech, and the perception studies examine pairs of phonemically identical sequences. For example, Spinelli et al. [13,14] compared 30 pairs of items like *l'affiche* 'the poster'/ *la fiche* 'the sheet' and *l'allocution* 'the (short) speech'//*la locution* 'the phrase, idiom', which differ systematically in the alignment of the AP-initial boundary rise (APR). In read speech, the APR consistently begins at the beginning of a content word; it therefore begins later in *la locution*, since the noun *locution* is preceded by a clitic function word (the definite article *la*) (See Figure 1). This pattern is quite robust; we have found it for all of the speakers in our various production studies, without exception.[1]

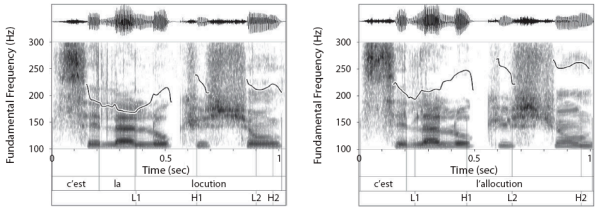

Figure 1: *Illustration of tonal alignment contrast in the pair* la location/ l'allocation *(from [14])*

This F0 rise, however, is more common in certain speaking styles (see, for example, [5]), including a "didactic" style and the style of newsreaders, and the extent to which it is available to listeners as a cue to word segmentation and lexical access in spontaneous speech is unclear.

The goals of this study were twofold: 1. to investigate whether the intonational patterns observed in read speech are also present in spontaneous speech and therefore available to listeners as cues to word segmentation and lexical access 2. to test two measures of characterizing intonation patterns using automatically extracted F0 and time values. Developing such automatic methods is important, since hand annotation is labor-intensive and often prohibitively expensive in time and money, particularly for large corpora of many hours of speech.

2. Methods

We analyzed the Corpus of Interactional Data (the CID), a corpus of conversational French speech. The corpus contains

[1] Spinelli et al. (2007) [13] used stimuli based on the read speech of a single speaker. Of the 30 critical pairs in that study, 29 were produced with the expected tonal alignment pattern. These 29 pairs were used in Spinelli et al. (2010) [14].

eight hours of audio/video recordings of conversations between eight pairs of speakers, yielding a total of 16 hours of speech [4]. The CID is fully or partially annotated on a number of levels, including: lexical, segmental, syntactic (part of speech), discourse, and gestural. The segmental annotation was performed using forced alignment and the lexical and syntactic annotation using a morpho-syntactic analyzer (for details, see [4]). Only a very small portion of the corpus is prosodically annotated, and this partial analysis was performed by hand.

Using information from the syllabic and syntactic annotation tiers, we searched the CID to find sequences comparable to the materials in our studies of read speech. We defined two types consisting of two-syllable sequences (s1, s2):

Type 1:
s1: determiner[1]
s2: first syllable of a noun or adjective[2]
e.g., *la locution, la chaussure, un moment*

Type 2:
s1, s2: first two syllables of a noun preceded by a determiner
e.g., *l'allocution, (la) chaussure, (un) moment*

For Type 2, we included only nouns and adjectives of at least two syllables.

The search returned a total of 7443 pairs. Syllables shorter than 72 ms or longer than 400 ms were excluded from the analyses, since these values are often due to alignment errors. Syllables whose median F0 appeared to be an outlier (men: < 60 Hz and > 230 Hz, women: < 150 Hz and > 300 Hz) were also excluded. This left 4724 pairs (3045 Type 1, 1679 Type 2) for the analyses.

For each pair, we extracted time points for the boundaries of s1 and s2, as well as F0 values, and calculated two types of measures of the F0 differences associated with the presence or absence of an intonational rise. As in our earlier studies, we attempted to quantify the F0 differences between conditions while avoiding the problems posed by segmental perturbations (e.g., due to the presence of voiceless obstruents). For the two-syllable target region, we calculated 1. median F0 for each target syllable (s1, s2), which takes into account only the voiced portion of the target region, and 2. normalized *Tonal Center of Gravity* (TCoG). The TCoG is a measure of the ''the overall distribution of the 'mass' or bulk of raised F0 in both time and frequency space'' [2].

3. Results

3.1. Automatic extraction of intonation patterns

3.1.1. Median F0

If AP boundary rises (APRs) are present, we would expect the difference in median F0 from s1 to s2 to be greater in Type 1 sequences (determiner+noun) than in Type 2 sequences (first two content-word syllables). As we have noted earlier, the Spinelli et al. (2007, 2010) [13,14] stimuli are characterized by clear and consistent AP boundary rises (APRs) beginning at content word-initial syllables. These stimuli were based on the speech of a single speaker reading targets (e.g, *la locution* (Type 1), *l'allocution* (Type 2)) embedded in a carrier phrase. A comparison between F0 median values of the first two syllables revealed the expected difference by sequence type (Type 1: 182 Hz vs. 216 Hz, Δ 34 Hz, Type 2: 200 Hz vs. 222 Hz, Δ 22 Hz).

A closer examination, however, revealed that the effect was sensitive to segmental composition. When a French accentual phrase begins with a clitic function word (such as a determiner), there is often a clear inflection point in the F0 curve at the function-word/content-word boundary (e.g., between *la* and *locution* and between *c'est* and *l'allocution* in Figure 1). This inflection point is clear when there are voiced consonants at the critical region (for example, the liquid [l]), but is not present (at least physically) when vocal fold vibration is interrupted by voiceless segments. In the Spinelli et al. (2007, 2010) [13,14] materials, the first syllable was always [la], either the feminine singular of the definite article *la* (*la locution, la fiche*) or an elided form of the definite article and the vowel of a vowel-initial noun (*l'allocution, l'affiche*). The segmental composition of the second syllable (s2) therefore varied, and it sometimes contained a voiced onset ([l] for s2 of *la locution/l'allocution*), sometimes a voiceless onset ([f] for the s2 of *la fiche, l'affiche*). Our analyses revealed that the difference in F0 median was significant only for items with a voiceless s2 onset

A linear mixed model was fitted, with the ratio s2/s1 of the medians as dependent variable, group (Type1, Type2) and voicing (voiced, voiceless) as predictors. A random intercept was added to account for the variability of the 29 pairs (e.g., *l'allocution*/*la locution*) across the 58 observations. The difference between the Type 1 ratio (1.22) and the Type 2 (1.144) ratio was significant (SE = 0.016, t = –4.7, $p < 0.0001$, Type 1: 182 Hz vs. 222 Hz, Δ 40 Hz, Type 2: 200 Hz vs. 224 Hz, Δ 24 Hz). However, this difference was not significant for those with a voiced s2 onset (Type 1 ratio (1.146) and Type 2 ratio (1.116) (SE = 0.018, t = –1.29, p = 0.2, Type 1: 182 Hz vs. 205 Hz, Δ 23 Hz, Type 2: 195 Hz vs. 222 Hz, Δ 27 Hz). Note that there is a significant difference between sequence type when the vowel alone contributes to the median F0 (when there are voiceless s2 onsets).[3]

A similar model was fitted on the CID with random intercepts to account for the variability of the 16 speakers and the 2089 phonetic sequences represented in the s1s2 sequences across the 4724 observations. The results show a similar pattern, although the differences are much smaller and must be interpreted with caution. For items with voiceless s2 onsets, the F0 median is significantly different, in the expected direction (the Type 2 ratio of the medians is significantly smaller than the Type2, ß = –0.043, SE = 0.0077, t = –5.66, $p < 0.0001$, Type 1: 178 Hz vs. 185 Hz, Δ 7, Type 2: 181 Hz vs. 185 Hz, Δ 4). For items with voiced s2 onsets, there was a significant difference between types, but in the opposite direction (ß = 0.035, SE = 0.0097, t = 3.7, $p < 0.0001$).

3.1.2. Tonal Center of Gravity

We expected that the Tonal Center of Gravity measure would be able to capture the intonational differences present in the

[1] The determiners were all monosyllables (*le, la, les, un, une, des, ma, mes, ta, tes, sa, ses, nos, vos…*).
[2] This was usually a noun, since French noun phrases are typically left-headed.

[3] Spinelli et al. 2007 [13] quantified F0 differences by measuring F0 at vowel midpoints. We performed additional analyses on the stimuli from that study.

Spinelli et al. (2007, 2010) [13, 14] materials and in particular that the measure would be robust in the face of segmental perturbations from voiceless segments. The TCoG was developed to address a range of difficulties for models of intonation based on tonal targets or turning points (TPs), including the interruption of the F0 by segmental perturbation: "Among the problems confronting a purely TP-based approach to intonational phonetics and phonology are the following: (1) Crucial TPs are frequently missing from the F0 curve owing to, e.g., intervals of voicelessness or irregular phonation" [2]. The TCoG "focus[es] not on the onsets and offsets of pitch movements, but rather on the overall distribution of the "mass" or bulk of raised F0 in both time and frequency space" [2].

We calculated the TCoG according to the formula in (1) (from [2]).

$$T_{cog} = \frac{\sum_i F0_i t_i}{\sum_i F0_i} \quad (1)$$

The time values obtained were normalized to 1, i.e., they were expressed as a proportion of the duration of the target two-syllable sequence. We predicted that the normalized TCoG (NTCoG) would be greater in Type 1 (determiner+noun) than in Type 2 (first two content-word syllables), corresponding to an F0 rise starting at the boundary between these two syllables (e.g., farther to the right).

We expected that the NTCoG measure would allow us to abstract away from perturbations of the F0 curve caused by voiceless segments (for which F0 is set to 0, in order to respect the definition of center of gravity). Our analysis of the Spinelli et al. 2007 materials, however, revealed no differences in NTCoG across sequence types (Type 1 vs. Type 2), even considering items with voiceless s2 onsets and voiced s2 onsets separately (for voiceless onsets, ß = −0.0018, SE = 0.007, t = −0.26, p = 0.79; for voiced onsets, ß = 0.0039, SE = 0.006, t = 0.49, p = 0.78).

In addition, the results show that the TCoG is sensitive to segmental composition. Within a sequence type, the presence of a voiceless s2 onset shifts the NTCoG to the left (ß = −0.035, SE = 0.014, t = 2.47, p < 0.05). We can illustrate why this should be the case with the simple example of a $C_1V_1.C_2V_2$ sequence, where C_1 is voiced and C_2 is either voiced or voiceless. All other things being equal, the first syllable (s1) C_1V_1, which is voiced throughout, will necessarily have a greater "mass" of F0 than C_2V_2 when C_2 is voiceless, since the F0 mass of the voiceless C_2 is zero. This shifts the TCoG to the left.

The CID corpus offered us the opportunity to examine the sensitivity of TCoG to segmental composition, specifically the influence of voiced versus voiceless segments, using a large corpus with a much greater variety in segmental composition (e.g., first syllables other than [la]). The results indicate that the measure is sensitive to the presence or absence of voiceless segments. Given the predominance of open syllables (e.g., CV, CCV) in French, we expect the presence of voiceless consonants in syllable onsets to shift the NTCoG to the right.

A linear mixed model was fitted, with group (Type 1, Type 2) and voicing as predictors and the NTCoG as the dependent variable. A random intercept was added to account for the variability across the 2089 phonetic sequences in the 4724 observations. The results show a very small but significant difference (ß = –0.011, SE = 0.0044, t = –2.48, p < 0.05) in the expected direction. The NTCoG for Type 2 is slightly smaller than for Type 1 (0.502 vs. 0.513), corresponding to a shift of only about 3 ms in two-syllable target sequences with a mean duration of about 300 ms. As expected, for items with voiceless syllable onsets, there was a strong shift to the right for Type 2 (ß = 0.07, SE = 0.003, t = 20, p < 0.0001). The comparison between Type 1 and Type 2 sequences with voiceless s2 onsets is complicated, however, by the fact that the many of most frequent determiners of French (s1 in Type 1 sequences) contain voiced segments (e.g., all the forms of the definite and indefinite articles: *le, la, les, un, une, des*).

4. Discussion and Conclusions

The study had two main goals: 1. to investigate whether the intonational patterns observed in read speech are also present in spontaneous, conversational speech, 2. to test two measures of characterizing intonation patterns using automatically extracted F0 values.

Each of the two measures tested proved problematic. The F0 median successfully quantified the intonational differences between the two sequence types in the Spinelli et al. (2007, 2010) [13,14] read speech materials, but only for items with a particular segmental composition. The normalized Tonal Center of Gravity (NTCoG) also failed to fully capture the intonational differences between sequence types and was sensitive to segmental composition in ways unpredicted by the hypothesis. Recent work on German and Italian also calls into question certain predictions of the TCoG hypothesis and offers another perception-based explanation of intonational contrast [7].

Turning to the comparison between speaking styles, the results provide no conclusive evidence that the accentual phrase boundary rises (APRs) found in careful, read speech are reliably present in the spontaneous speech of the CID corpus, although durational differences are present. There were significant differences across sequence types only for certain limited cases.

Whether or not further study shows that AP-initial boundary rises are present in spontaneous speech *of this type*, however, this intonational pattern may still be reliably present in other styles of spontaneous speech and therefore available to listeners as a cue to word segmentation and lexical access. Beckman (1997) [3] identifies ten types of "spontaneous speech" (see also discussion in [20]). So, for example, early rises may be present in spontaneous, but careful speech or in speech directed at young children.

Our analyses of the CID were conducted on the totality of the corpus. However, speaking style in conversational speech is not homogeneous, and different styles may be present even within a single dialog or conversation. For example, factors such as the organization of the discourse and interactions between interlocuters may create alternating stretches of clear, even hyper-articulated speech and stretches of extremely reduced, hypo-articulated speech. We plan to conduct further analyses to automatically search for stretches of clear, careful speech in the corpus. This will allow us to examine the hypothesis that prosodic cues for lexical segmentation may be more reliably present in spontaneous, conversational speech, when the speech is careful, clear and slower, than when the speech is faster and more subject to segmental and prosodic reduction.

Segmental durations are quite short for most of the segments of the CID [11]. To evaluate our hypothesis, we plan to identify stretches of speech with longer segmental durations

(e.g., greater than 150 ms). If prosodic cues for word segmentation are used in conversational speech, we expect that they will appear in these stretches.

We note that other types of cues to word segmentation, such as those based on segmental transitional probabilities, may be less reliable in casual than in careful speech. For example, the massive segmental reduction found in casual speech (see [11] on reduction in the CID) may give rise to word-initial segment sequences unattested in citation forms (e.g., [pti] for *petit* 'small'). These style-dependent differences are not unexpected, and the context-dependent weighting of cues to word segmentation and lexical access is included in a recent model [10].

5. References

[1] Astésano, C., Bard E.G. and Turk, A. Functions of the French initial accent: A preliminary study. Speech Prosody Proc., Aix-en-Provence, 139–142, 2002.

[2] Barnes, J., Veilleux, N., Brugos, A. and Shattuck-Hufnagel, S. The effect of global F0 contour shape on the perception of tonal timing contrasts in American English intonation. Speech Prosody, Chicago, 2010.

[3] Beckman, M. E. A typology of spontaneous speech, in Y. Sagisaka, N. Campbell & N. Higuchi [Eds], Computing prosody: Computational models for processing spontaneous speech, 7–26, New York: Springer Verlag, 1997.

[4] Bertrand, R., Blache, P., Espesser, R., Ferré, G. Meunier, C. Priego-Valverde, B. and Rauzy, S. Le CID – corpus of interactional data. *Traitement Automatique des Langues* 49: 105–134, 2008.

[5] Di Cristo, A. Vers une modélisation de l'accentuation du français : première partie. *French Language Studies* 9: 143–179, 1999.

[6] Di Cristo, A. Vers une modélisation de l'accentuation du français : seconde partie. *French Language Studies* 10, 27–44, 2000.

[7] D'Imperio, M., Gili Fivela B. and Niebuhr, O. Alignment perception of high intonational plateaux in Italian and German. Speech Prosody Proc., Chicago, 2010.

[8] Jun, S.-A. and Fougeron, C. Realizations of Accentual Phrase in French. *Probus* 14, 147–172, 2002.

[9] Kim, S. and Cho, T. The use of phrase-level prosodic information in lexical segmentation: Evidence from word-spotting experiments in Korean. *Journal of the Acoustical Society of America* 125: 3373–3386, 2009.

[10] Mattys, S. L. and Melhorn, J.F. Sentential, lexical, and acoustic effects on the perception of word boundaries. *Journal of the Acoustical Society of Americ*a 122: 554–567, 2007.

[11] Meunier, C. and Espesser, R. Vowel reduction in conversational speech in French: The role of lexical factors. *Journal of Phonetics*, in press.

[12] Michelas, A. and D'Imperio, M. Accentual Phrase boundaries and lexical access in French. Speech Prosody Proc., Chicago, 2010.

[13] Spinelli, E., Welby, P. and Schaegis, A.-L. Fine-grained access to targets and competitors in phonemically identical spoken sequences: The case of French elision. *Language and Cognitive Processes* 22: 828–859, 2007.

[14] Spinelli, E., Grimault, N., Meunier, F. and Welby, P. Spinelli, E., Grimault, N., Meunier, F. and Welby, P. An intonational cue to word segmentation in phonemically identical sequences. *Attention Perception et Psychophysics* 72: 775–787, 2010.

[15] Vaissière, J. Langues, prosodies et syntaxe. *Traitement Automatique des Langues* 38: 53–82, 1997.

[16] Warner, N., Otake, T. and Arai, T. Intonational structure as a word-boundary cue in Tokyo Japanese. *Language and Speech* 53: 107–131, 2010.

[17] Welby, P. The slaying of Lady Mondegreen, being a study of French tonal association and alignment and their role in speech segmentation. Ph.D. dissertation, The Ohio State University, 2003.

[18] Welby, P. French intonational structure: Evidence from tonal alignment. *Journal of Phonetics* 34: 343–371, 2006.

[19] Welby, P. The role of early fundamental frequency rises and elbows in French word segmentation. *Speech Communication* 49: 28–48, 2007.

[20] Xu, Y. In defense of lab speech. *Journal of Phonetics* 38: 329–336, 2010.