Investigation of lexical f_0 and duration patterns in French using large broadcast news speech corpora

Rena Nemoto[1], *Martine Adda-Decker*[1], *Jacques Durand*[2]

[1]LIMSI-CNRS, Orsay, France
[2]CLLE-ERSS/CNRS & Université de Toulouse-Le Mirail, France
{nemoto,madda}@limsi.fr, jacques.durand@univ-tlse2.fr

Abstract

This work aims at improving our knowledge of links between prosody and pronunciation variants in French. An original methodology is proposed to study prosodic regularities of French words via average f_0 profiles, by making use of automatic processing and 13 hours of broadcast news speech. Investigated influential factors include word syllable length, duration, word-final schwa, parts of speech. The following questions are addressed: can specific lexical f_0 profiles be measured automatically using large corpora? If so, how do they vary with respect to the cited influential factors? Results confirm the known tendency of word-final syllable accentuation. They also highlight some word-initial accentuation. Higher average f_0 profiles are measured for increasing segment durations (locally decreasing speaking rate), but also for words ending with schwas. Future studies include phrase boundary annotation and the extension to a larger variety of speaking styles and languages.

Index Terms: f_0 profile, syllabic word length, lexical duration, word-final schwa, French

1. Introduction

Taking a long-term perspective, this work aims at improving the acoustic modeling capacities in automatic speech recognition by increasing our knowledge of links between prosody and pronunciation variants. To this aim the proposed study investigates prosodic regularities of French words in large speech corpora.

Concerning human speech processing, a large body of works have addressed the question of whether and how word boundaries may be inferred from the acoustic signal by human listeners. A review of the literature on human word segmentation reveals two main tendencies: (i) the word segmentation problem can be – at least partly – solved by distributional properties of the language [10, 15, 17], (ii) the word segmentation problem takes benefit from acoustic cues among which most importantly prosodic information [2, 5, 14]. Automatic speech recognition (ASR) systems tend to hypothesize word boundaries in continuous speech using word and word co-occurrence information, rather than specific acoustic cues. ASR systems can then be viewed as supporters of the first trend, relying on distributional cues. These come from the lexical level, rather than from prelexical levels in psycholinguistic studies, as ASR systems get a priori knowledge of a language's lexicon. However, the word segmentation problem remains tricky due to combinatorial complexity. In [4], Cutler et al. review the need of prosody for word boundary location as part of prelexical processing. In particular, the importance of relative syllable durations was highlighted for English word boundary location, but also for French [16] homophone phrases such as *le couplet complet* (parts of speech: `det noun adj`) vs *le couple est complet* (`det noun verb adj`) /ləkuplɛkɔ̃plɛ/. Such multiword homophones, where phonotactic distributional cues are canceled out, demonstrate the importance of prosodic cues for word segmentation. What are these cues in French and can they contribute to predict word boundaries and to explain pronunciation variation?

By relying on raw f_0 and segment duration measurements, as provided by automatic speech alignments, average f_0 profiles of French words are computed. The questions addressed are the following: can specific f_0 profiles for French words be measured automatically using large corpora? If so, how do they vary with respect to influential factors, such as word syllable length, the presence of final schwas, vocalic or syllabic durations or part of speech categories? Concerning our knowledge of links between prosody and pronunciation, it has been shown that segment duration influences vowel timber and quality [9, 13], that syllable coda consonants are more prone to deletion than syllable initial consonants etc., that fluent speech may give rise to "speech reductions" which are specific to natural, native speech, but difficult to reproduce by non-native speakers, and potentially harmful for automatic speech recognition devices. The aim of this study is then to produce empirical evidence concerning the raised questions, in order to contribute to our knowledge of prosodic realizations in French words, their potential to contribute to the word segmentation and the pronunciation variation problems.

Section 2 presents the speech corpus and the methodology to extract and organize measurements. Section 3 presents f_0 profiles as a function of influential factors and section 4 provides additional information concerning intervocalic f_0 and duration measurements. Conclusions are presented in section 5.

2. Corpus and Methodology

2.1. Corpus

This study makes use of 13 hours of male speech from the manually transcribed French TECHNOLANGUE-ESTER corpus [7] (news from different Francophone radio stations). The genre of speech mainly consists of broadcast news presented by professional speakers at a normal to sustained pace with only few stops or breaks. The speaking style can be qualified as globally neutral, with functions of enunciation and demarcation prevailing over lively expression and emotions. The examined data include 165k word tokens and 14k word types. Table 1 shows the corpus composition according to mono-/polysyllabic words.

2.2. Methodology

Concerning French prosody, many authors have noticed the correlation between accentuation (final and initial), lengthening on

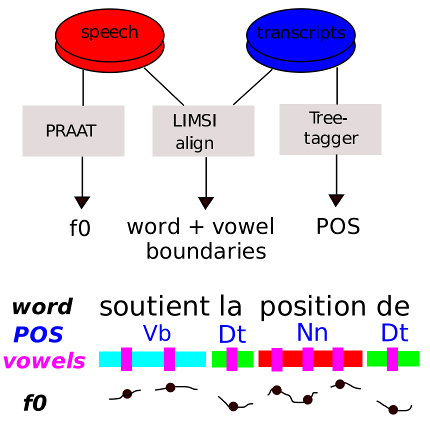

Figure 1: *Automatic processing steps and annotation levels: each vowel is tagged by an average f_0 value and its duration, by its rank within the word, by lexical and POS information.*

Table 1: *Quantitative corpus description w.r.t. word tokens of syllable length n from* 0 *to* 4*. Counts are separated for words w/wo realized final schwa(top/bottom). Syll.class n_s states n: the number of full syllables; s: presence/absence of final schwa.*

n	*Syll.class* n_s	*#Words*	*Examples*
0	0_0	12578	l'; d'; de
1	1_0	72249	vingt; reste
2	2_0	36027	beaucoup; journal
3	3_0	15994	notamment; militaire
4	4_0	6053	présidentielle

n	*Syll.class*	*#Words+ /ə/*	*Examples*
0	0_1	12295	de; le; que
1	1_1	3918	reste; test
2	2_1	2087	ministre
3	3_1	698	véritable
4	4_1	174	nationalistes

prosodic phrase boundaries [1, 6, 11, 12, 20]. In the following, we do not try to locate phrase boundaries, however we propose contrastive measurements on subsets with increasing proportions of potential prosodic phrase boundaries. Acoustic correlates, namely f_0 and durations are examined with respect to supposed influential factors: word length expressed in number of syllables, presence or absence of word-final schwa, POS, speaking rate. Figure 1 gives a schematic overview of the processing steps on the investigated data.

F_0 measurements: Fundamental frequency (f_0) values were measured every 5 ms using the standard settings of Praat [3].

Lexical and phonemic alignment: The audio corpus was automatically aligned by the LIMSI speech recognition system [8] producing word and phoneme segmentations. The pronunciation dictionary was tuned to propose optional word-final schwas for pronunciations ending in a consonant.

Word syllable length; Syllable length class: Each word token was annotated by its *syllable length*, corresponding to the number of full syllables in its aligned pronunciation. Word-final schwas did not count for the syllable length, however, they were used to tag words into specific subsets. The word *reste* ('rest') with pronunciation [ʁɛst] was of syllable length 1 with no word-final schwa, and was tagged as belonging to the *syllable length class* 1_0. The same word pronounced [ʁɛstə] goes to the *syllable length class* 1_1 (cf. *syll.class* in Table 1). Words of the same syllable class are merged to compute average f_0 profiles.

Part Of Speech (POS) tagging: To examine the impact of syntactic classes on f_0 realizations, POS were semi-automatically tagged using a French version of TREETAGGER [18].

F_0 values, f_0 profiles: Each aligned vowel segment with voicing ratio over 70% was given an *f_0 value* corresponding to the average of all its individual 5 ms measurements (different ways of computing f_0 values on more or less restricted central parts were tested without major changes on final profiles). The f_0 values in Hz were converted to semitones (ST), with 120 Hz as baseline frequency (120 Hz is often considered as the average male voice height [19] and was actually close to the average f_0 of our corpus). Only words with all their vowels passing the voicing criterion were kept for further investigations. This selection aimed at reducing the impact of erroneous measurements, due to combined alignment and/or f_0 extraction errors. The *f_0 profile* of a word was then defined as a schematic f_0 contour connecting the f_0 values of the different vowels (of increasing syllabic rank) of this word. Similarly, for a given word class (e.g. *syll.class* in Table 1), the *f_0 profile* could be defined as connecting average f_0 values of increasing rank, where the average f_0 value of a given syllabic rank was computed over all the vowels of this rank in the considered word class. For example, given the 2_0 class of bisyllabic words without final schwa, the corresponding f_0 profile was computed as the contour connecting the average f_0 value of the rank 1 vowels (first syllable) to the average f_0 value of the rank 2 vowels. Further word subsets are added using POS information.

3. F_0 profile results

In the following, f_0 profiles are presented according to the introduced n_s syllable length classes. First we present profiles for lexical words as opposed to grammatical words. The rationale is to empirically confirm that grammatical words tend to remain unstressed which should then result in comparatively lower f_0 profiles. Further profiles show differences according to absence or presence of final schwa. Then we focus on nouns and the (Det - Noun) phrases.

As French tends to produce word-final accentuation, the graphical displays of the f_0 profiles of increasing syllabic length were right-justified: the first syllable of monosyllabic words, the second syllable of bisyllabic words etc. are displayed at the final n-th position of the longest n-syllabic words.

3.1. Lexical vs. grammatical words

Most occurrences of grammatical words in French are mono- or bisyllabic, whereas lexical words are frequently polysyllabic. Due to minimum word frequency criteria, all profiles are limited to at most 4-syllabic lexical and bisyllabic grammatical words. Figure 2 shows the corresponding f_0 profiles. **Lexical word** profiles show that

(i) Mean f_0 is much higher for the final syllable n than for all preceding syllables (Δ1-3 ST more).

(ii) For trisyllables or more, the f_0 difference between two consecutive vowels is maximal between penultimate and final vowels (Δ2-3 ST). The corresponding delta increases with word syllable length.

(iii) Mean f_0 of monosyllables is at least as high as the ones of the final syllable of polysyllabic words.

(iv) Evidence of initial word accent remains weak on mean f_0 contours.

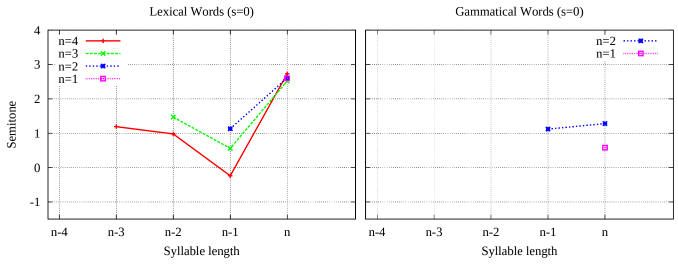

Figure 2: *Mean f_0 profiles of n-syllabic words.* **Left:** *Lexical words (n=1-4, s=0)* **Right:** *Grammatical words (n=1-2, s=0).*

For grammatical words, relatively low f_0 values can be observed. Average f_0 contours feature flatter curves than the lexical word ones where relatively steep falls can be observed on the penultimate syllable.

3.2. Lexical words w/wo word-final schwa

Word-final schwas change the rhythmic pattern of speech through the addition of an unstressed syllable. Observation counts (see Table 1) show that only a limited amount of word tokens got realized word-final schwas in standard French broadcast news speech. Figure 3 displays f_0 profiles both for lexical words without schwas (left, the same as Figure 2 left) and for lexical words with final schwas (right). This comparison shows that:

(i) A final schwa triggers a slight increase of mean f_0, in particular for the final full syllable n.

(ii) The difference between the final syllable n and the final schwa, corresponding to 2-3 ST, is greater than the delta between final and penultimate syllables.

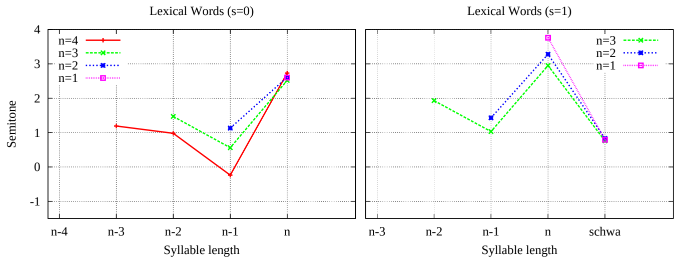

Figure 3: *Mean f_0 profiles of n-syllabic words.* **Left:** *Lexical words (n=1-4, s=0)* **Right:** *Lexical words (n=1-3) with word-final schwa (s=1).*

3.3. Short vs long duration

To investigate the impact of duration on f_0 profiles we separated lexical words in fast rate words and slow rate words by filtering according to vocalic duration on all but the final syllable. The words with all such vocalic durations lower than 75 ms are considered as the (locally) fast rate words, whereas the remaining ones correspond to the (locally) slow rate speech. Figure 4 shows the corresponding results. Proportions of investigated words between fast rate words and slow rate words in Figure 4 are; **fast** rate vs. **slow** rate, bisyllabics: 68% vs. 32%, trisyllabics: 56% vs. 45%, 4-syllabics: 52% vs. 48%. If we look into our speech corpus, 60% of vowels belong to lower than 75ms and 40% of vowels to slow rate speech. Short duration words (left figure) display much lower f_0 profiles as compared to longer lasting words (right figure). This result suggest that speech rate deceleration correlates with a global upward f_0 trend of the corresponding words. Further studies need to more specifically address speech deceleration in sentence-final positions, where f_0 is expected to drop.

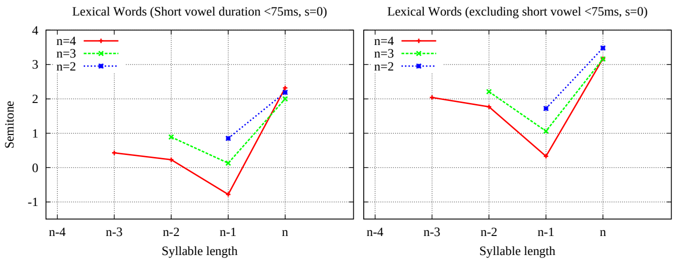

Figure 4: *Mean f_0 profiles of n-syllabic lexical words (n=1-4, s=0) as a function of duration.* **Left:** *Short vocalic durations (<75ms, except for final vowels)* **Right:** *All but the short ones.*

3.4. Noun phrase

To address the question of f_0 profiles across word boundaries, we examined mean f_0 profiles for noun phrases. The calculation was limited to the determiner noun bigram. Are mean f_0 profiles of an n length noun phrases different from an n length noun? Figure 5 (left) shows the mean f_0 profiles of Noun words (31k occ.), very similar to Figure 2 (left). The right figure exhibits the mean f_0 profiles of noun phrases (13k occ.).

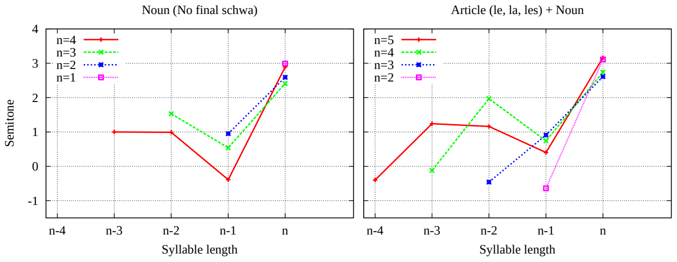

Figure 5: *Mean f_0 profiles for n-syllabic length.* **Left:** *Nouns (n=1-4)* **Right:** Determiner noun *phrases (n=2-5).*

For a given syllabic length n, profiles are quite different depending they represent just a noun or a det-noun sequence. Whereas for a noun the profile drops from the initial starting value to a minimum on the penultimate syllable to raise to an absolute maximum on the final syllable, det-noun phrases first start with a low f_0 value on det with a first rise to the noun-initial syllable. This information may be of help to locate word boundaries and to disambiguate homophones such as déblocage ('unblocking') and des blocages ('blockings').

4. Inter-vocalic measurements

This section presents **intervocalic duration** statistics with the double motivation. Firstly we would like to examine these durations with regard to the previous f_0 profiles. Second motivation is, taking our long term perspective, to spot subwords particularly prone to temporal reduction and possibly to shorter pronunciation variants. For a given vowel of rank n, its inter-

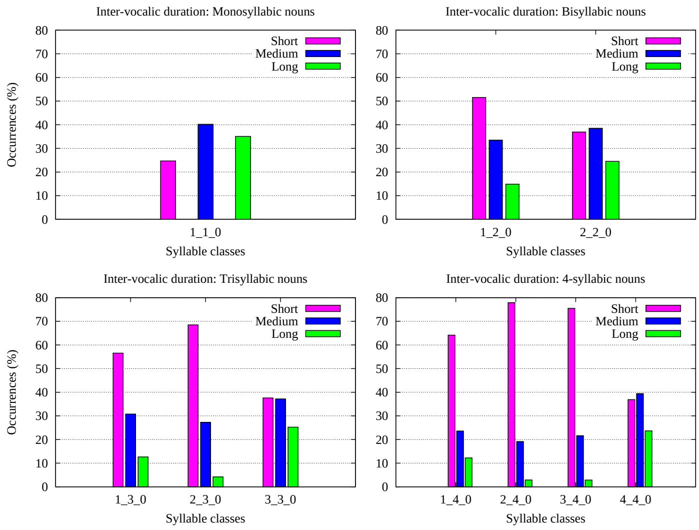

Figure 6: *Histograms of inter-vocalic durations for nouns w.r.t. vowel rank (Table 1).* **Top Left:** *monosyllabics.* **Top Right:** *bisyllabics.* **Bottom Left:** *3-syllabics.* **Bottom Right:** *4-syllabics.*

vocalic duration measures the time span between the centers of the given vowel and its preceding vowel. For first syllables, the preceding vowel corresponds to the last vowel of the preceding word. The intervocalic duration can be seen as an approximation of syllabic duration. Similar measurements have been carried out for intervocalic f_0 differences. These show that noun word final f_0 fall patterns (< -1 ST) are about 20%, whereas 70% correspond to final rise (> 1 ST). Figure 6 shows the statistics for mono-, bi-, tri- and 4-syllabic nouns, using the following duration classes:

Short	Medium	Long
$[30ms - 155ms]$	$[160ms - 220ms]$	$[225ms - [$

A comparison of the 4 figures (Figure 6) reveals that the proportions of shorter intervocalic duration class (Short) increase with syllabic word length. However, as expected, the final vowel position yields much lower rates for all polysyllabic configurations. Results confirm that longer words tend to be uttered more rapidly with the internal syllables particularly prone to temporal shortening. These subword parts are to be considered in future pronunciation dictionary design for ASR systems.

5. Conclusions

Lexical f_0 and duration patterns in French were investigated using 13 hours of broadcast news audio (165 000 words) of male speakers. An original methodology is proposed to study prosodic regularities of French words via average f_0 profiles. The presented methodology makes use of time-aligned phonemic and lexical transcriptions, as well as of prosodic and POS annotations to compare average f_0 profiles according to word classes of given syllabic length, word final-schwa, duration and phrases. The average lexical word contour exhibits a final rise concurrent with a minimum f_0 on the penultimate syllable, which tends to be reinforced with syllabic word length. Average f_0 profiles tend to raise in presence of final schwa and for longer syllabic durations. These findings contribute to our knowledge of links between prosody and pronunciation variants in French. Future studies will include phrase and sentence boundary annotations as well as the extension to a larger variety of speaking styles and languages. The achieved results tend to show that acoustic-prosodic features may be helpful to word boundary localization in French. Their successful implementation in future ASR systems remains a challenge.

6. Acknowledgements

This work was funded by a DIGITEO research cluster - *Région Ile-de-France* - doctoral grant and by the OSEO *Quaero* program.

7. References

[1] Adda-Decker, M., et al., “Contributions du traitement automatique de la parole à l’étude des voyelles orales du français”. In TAL Vol. 49, No 3. Phonétique et Phonologie, 13-46, 2008.

[2] Bagou, O., Fougeron, C., Frauenfelder, U. H., “Contribution of prosody to the segmentation and storage of "words" in the acquisition of a new mini-language”. In Proceedings of Speech Prosody, 59-62, Aix-en-Provence, France, 2002.

[3] Boersma, P., Weenink, D., “Praat: doing phonetics by computer [computer program]”, http://www.praat.org/, Tech. report, 2005.

[4] Cutler, A., Dahan, D., and Van Donselaar, W. “Prosody in the comprehension of spoken language: a literature review”. In Language and Speech, 40(2):141-201, 1997.

[5] Cutler, A. and Norris, D., “The role of strong syllables in segmentation for lexical access”. In Journal of Experimental Psychology: Human Perception and Performance, 14:113-121, 1988.

[6] Fougeron, C., Jun, S. A., “Rate Effects on French Intonation: Prosodic Organization and Phonetic Realization”. J. of Phonetics, 26:45-69, 1998.

[7] Galliano, S., et al., “The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News”. In Proc. Interspeech, Lisbonne, 2005.

[8] Gauvain, J.-L., et al., “Where Are We In Transcribing French Broadcast News?”. In Proceedings Interspeech, Lisbonne, 2005.

[9] Gendrot, C. and Adda-Decker, M., “Impact of duration and vowel inventory size on formant values of oral vowels: an automated formant analysis from eight languages”.

[10] Harris, Z., “From phoneme to morpheme”. In Language, 31:190-222, 1955.

[11] Hirst, D., Di Cristo, A., Intonation Systems : A Survey of 20 Languages, Cambridge University Press, Cambridge, 1998.

[12] Lacheret-Dujour, A. and Beaugendre, F., La Prosodie du Français, CNRS Éditions, Paris, 1999.

[13] Lindblom, B., “Spectrogaphic study of vowel reduction”, Journal of the Acoustical Society of America, Vol. 35, pp 1773-1781, 1963.

[14] Mattys, S. L., Jusczyk, P. W., Luce, P. A., and Morgan, J. L., “Phonotactic and prosodic effects on word segmentation in infants”. In Cognitive Psychology, 38(4):465–494, 1999.

[15] McQueen, J. M., “Segmentation of continuous speech using phonotactics”. Journal of Memory & Language, 39:21-46, 1998.

[16] Rietveld, A. C. M., “Word boundaries in the French language”. In Language & Speech, 23:289-296, 1980.

[17] Saffran, J. R., Newport, E. L., and Aslin, R. N., “Word segmentation: The role of distributional cues”. In Journal of Memory & Language, 35:606-621, 1996.

[18] Schmid, H., “Probabilistic Part-of-Speech Tagging Using Decision Trees”. In Proceedings of International Conference on New Methods in Language Processing, Manchester, 1994.

[19] t’Hart, J., “Differential sensitivity to pitch distance, particularly in speech”. In Journal of Acoustical Society of America, 69(3):811-821, 1981.

[20] Vaissière, J., “Rhythm, accentuation and final lengthening in French”. In Sundberg, J. et al. [ED], Music, Language, Speech and Brain, 108-121, 1991.