A NEW METHOD FOR SPEECH DELEXICALIZATION, AND ITS APPLICATION TO THE PERCEPTION OF FRENCH PROSODY

V. Pagel, N. Carbonell, Y. Laprie

CRIN-CNRS & INRIA Lorraine, BP 239, 54506 Vandœuvre-Lès-Nancy, France

ABSTRACT

This paper describes a perception experiment which aims at testing the ability of subjects to discriminate prosodic boundaries when provided with fundamental frequency and duration. Their task consists in listening to a high quality speech synthesizer producing nonsense speech that copies the prosodic parameters of a French corpus. Subjects have marked boundary syllables with a device that synchronizes sound and orthographic transcription. Results show that the delexicalization procedure is ecological, and that duration strongly influences the perception of fundamental frequency.

1. MOTIVATIONS

Traditional delexicalization procedures offer a way for studying the contribution of intonation and rhythm to the segmentation of speech into suprasegmental units. Thanks to recent advances in speech synthesis, it now becomes possible to obtain more natural speech stimuli than those produced by methods such as random-splicing [1], gated speech [2], rotated speech, reiterated speech, and many kinds of bandpass filterings [3].

In this paper, we propose a new method for delexicalization based on the free MBROLA [4] speech synthesizer. We have performed a perception experiment based on this method, in order to elicit how listeners interpret F0 contours and segment duration variations in terms of prosodic boundaries, with a view to defining robust acoustic cues for the automatic segmentation of continuous spontaneous speech into prosodic groups.

2. DELEXICALIZATION OF FRENCH

Our idea is to modify the phones in a sentence, as little as possible, so that it cannot be understood anymore. Changes are summarized in Tab. 1.

When possible, consonants have been transformed into dental consonants for obtaining a unique place of articulation. However, voiced and unvoiced features in the original speech stimulus are kept untouched, and semi-vowels and liquids are left untouched as well, which avoids having to modify original F0 values (that would be quite difficult, e.g. for /R/ whose allophones can be voiced or unvoiced).

Phones	Replaced by
/p/ /t/ /k/	/t/
/b/ /d/ /g/	/d/
/f/ /s/ /S/	/s/
/v/ /z/ /Z/	/z/
/m/ /n/	/n/
/N/ /R/ /l/ /H/ /w/ /j/	Unchanged

Table 1: Phone transformation (SAMPA alphabet)

Vowels bear the main modifications: they are replaced by the centralized vowel /E/, but alternate vowel /A/ is used to avoid /E/ /E/ sequences.

In order to control intrinsic phone durations we have computed Z-scores [5] over our corpus. When a phone *i* is replaced by another phone *j* the new duration is:

$$newduration_j = \sigma_j * \frac{duration_i - \mu_i}{\sigma_i} + \mu_j$$

where μ is the average duration for the phone considered and σ the standard deviation of the duration. This correction is particularly important in French for nasal vowels ($/\mathrm{a\tilde{}}/$, $/\mathrm{o\tilde{}}/$, $/\mathrm{i\tilde{}}/$) which are intrinsically much longer than oral vowels.

As the energy is not a parameter of the Mbrola speech synthesizer, we cannot handle phenomena such as spectral tilt. By chance, in French, energy is a weak correlate of accentuation [6].

3. THE "PROSODIC KARAOKE"

The "Karaoke" method that we now describe is the means we propose for subjects to point out acoustic events they perceive in the delexicalized speech. A syllabic transcription is written on a computer screen where the subject can select an interval between two syllables with the mouse, listen to it, and eventually mark syllables. Just like in a true Karaoke, a cursor follows the orthographic transcription as it is being played. This cursor is important as it is the link between the acoustic and visual representations of the sentence.

In our experiment, the subject is free to select any part and listen to it as often as he needs. He can also interrupt playback immediately when something catches his attention. Syllabification was

achieved by hand, and orthographic transcription was done so that reading aloud matches with the phonetic transcription. For example "La presse aujourd'hui, nous pose, vous pose deux questions" [SOUND A920S04.WAV] becomes on the screen "lè très è zèr duè nè tèz zè tèz dè tès tyè" [SOUND A441S01.WAV].

4. EXPERIMENT

The speech synthesizer allows us to modify the prosodic parameters of the delexicalized speech. There are 4 conditions:

1. **F+D**: original F0 and duration [SOUND A441S01.WAV]
2. **D**: constant F0, original duration [SOUND A441S02.WAV]
3. **FM**: original F0, constant duration [SOUND A441S03.WAV]
4. **FS**: same as **FM** but slower [SOUND A441S04.WAV]

In condition **F+D**, fundamental frequency is copied from the original sentence. Three F0 values per phone is enough for a good reproduction of the original intonation, so we divide each phone in three equal parts and take one average F0 value on each parts. Duration is also copied from the original, but as explained before, it has passed through the Z-score normalization so that, for example, a 85 ms long /a~/ in the original sentence becomes a 62 ms long /E/ in the delexicalized one.

In condition **D** the fundamental frequency is the average F0 of the speaker, i.e. 121 Hz. Though the pitch remains constant, rhythm is copied from the original sentence after the Z-scoring normalization.

In condition **FM** the fundamental frequency is taken from the original sentence, and duration of each segment is the mean duration of the class the segment belongs to ($duration_i = \mu_i$) . As a result, the rhythm sounds fairly constant, but as we use the mean duration, the overall speech rate is the one of unaccented syllables.

After preliminary experiments, condition **FS** was added because we noticed that the subjects found that **FM** sounded quite fast, giving the impression of speech without moments for breathing (and for the subjects, the impression of no time left to integrate pitch movements). We constantly lengthened durations by adding 40 percent of the standard deviation ($duration_i = \mu_i + \sigma_i$*0.4), as a result the perception of final accents is enhanced.

For all the conditions, silent pauses and breathings were systematically shortened to 80 ms, they were not completely suppressed to avoid artificial pitch movements which would have extended between voiced phones on each side of a pause.

4.1. The corpus

We are currently studying "controlled speech", i.e. radio news announcements and press reviews, with a view to exploit the results in the framework of automatic recognition of continuous speech. Our corpus corresponds to a nine minute segment of a radio press review, spoken by a single speaker. It was phonetically annotated by a phonetician, and prosodically coded in a previous experiment [7] by an expert on prosody, and by automatic devices (namely multi-layer perceptrons).

	F+D	D	FM	FS
F+D	1	0.86	0.52	0.66
D		1	0.47	0.64
FM			1	0.71
FS				1

Table 2: Correlation between *agreements* across conditions

For this experiment we selected six extracts from the corpus so that each extract constitutes a self-sufficient prosodic paragraph. Some of the extracts include prosodic configurations that systematically triggered errors during our previous experiments with MLPs.

4.2. Protocol

The first session concerns the first three conditions only: **F+D**, **D**, **FM**. Eighteen naive subjects with no known auditory deficiency or difficulty to use a computer mouse segmented 6 sentences (2 for each condition), which takes about 40 minutes, including a training phase. They were given instructions to find all the word group boundaries they could, with a minimum confidence. The instructor gave them an example of different word groupings in the same sentence. At the end of the experiment, a questionnaire allowed them to judge the difficulty of the different conditions.

During the second session, two weeks later, they segmented the 2 sentences corresponding to condition **D**, under condition **FS**.

5. RESULTS

We organized our experiments so that each syllable for each condition could be marked as being a boundary by 6 subjects. On average a subject put ten marks on each sentence, and sentences lengths range from 43 to 90 syllables. For a mark to be statistically significant with $p<0.001$, it must have been found by 5 or 6 subjects. Generally an agreement for 4 subjects gives $p<0.05$, therefore we will take it as being the lower value for considering that a mark is significant in our analysis. In the rest, we call "*agreement*" the number of marks borne by a syllable (maximum is 6).

The questionnaire after each experiment allows a first validation concerning the delexicalization procedure: nobody was able to recognize words in the delexicalized corpus. Second validation, *all* the **F+D** marks with an *agreement* greater than 3 are positioned on syntagmatic or discursive boundaries. We deduce that the subjects found themselves in conditions that share structure properties with real speech, and that they managed when multi layer perceptrons were trapped in a previous experiment [7].

5.1. Comparison of the conditions

As seen in Tab. 2, the marks of **F+D** and **D** conditions are very similar, which implies that subjects where mainly looking for durational cues. This high correlation finds an explanation in Tab. 3, which teach that a significant *agreement* implies that the syllable is most likely followed by a pause in the original sentence.

Moreover, it reveals that none of the 277 unmarked syllables for

Agreement	none	1	2	3	4	5	6
F+D	0/277	0/40	2/10	7/9	3/4	6/9	23/26
D	0/267	1/40	5/15	1/9	4/11	4/4	27/29
FM	9/245	8/64	7/31	9/20	2/6	5/8	1/1
FS	6/233	3/64	5/33	6/18	8/13	7/8	6/6

Table 3: Portion of marks followed by a pause vs *agreement*

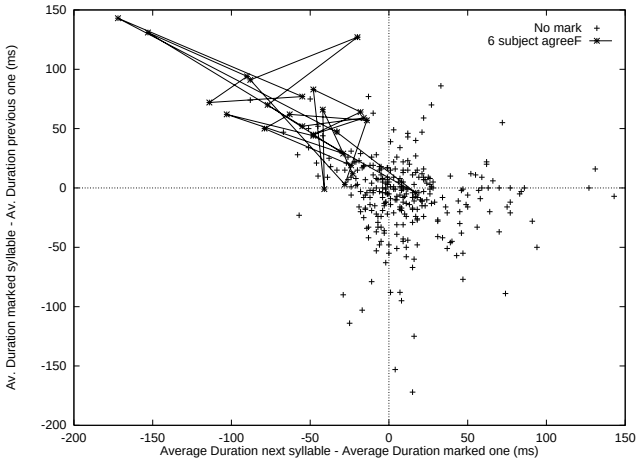

Figure 1: Delta duration of the marked syllable in the F+D condition

F+D were followed by a pause, which means that all the syllables preceding a pause bear at least one mark. All the pauses were also found in condition **D**, but this is not the case for **FM** and **FS**. Those two conditions show lower average *agreement*, with a slight advantage for **FS** over **FM**, and as a matter of fact, the subjects found them more difficult.

Fig. 1 illustrates how duration characterizes syllables marked in **F+D** condition. This figure represents the difference of average durations between the marked syllable and the next one on x axis, and the difference between the marked one and the previous one on y axis. The average duration is the syllable duration divided by its number of segments as in [8], to normalize against syllable lengths.

Fundamental frequency values of marked syllables are represented Fig. 2 for condition **F+D**. It appears that F0 extrema are not seeked by listeners when not associated with lengthening.

5.2. Intra-syllable analysis

In addition to this statistical study we tried to identify the acoustic cues on which the subjects' assessment of prosodic boundaries might be based.

Method
The six extracts from the corpus were synthesized from the original phonetic transcript, F0 contours and segment durations (using Mbrola). And the prosodic boundary marks set by a trained listener served as a reference perceptual prosodic segmentation. The expert also marked focus (or emphatic) accents during a second listening of the extracts.

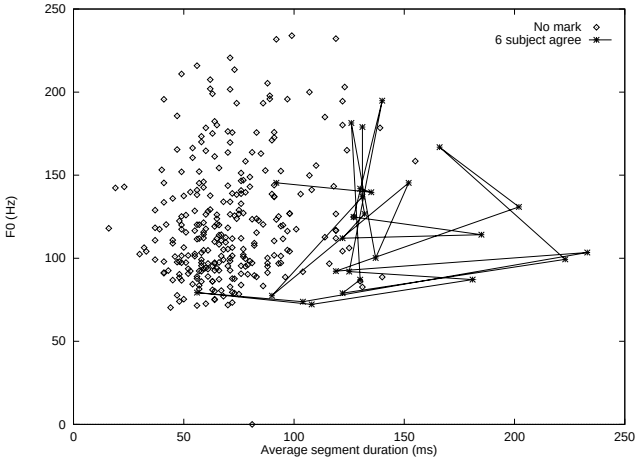

Figure 2: Average duration vs F0 of the marked syllable in **F+D**

The speech signal corresponding to these marks and the ones set by the subjects were then characterized in terms of F0 movements and syllabic duration variations, using the following features. We considered three types of syllabic lengthenings according to the phone(s) affected by the lengthening:

- all the phones in the syllable (*SL*),
- the vowel and subsequent sonorant(s) or syllabic nucleus (*VL*),
- the initial consonant (*CL*),

since French is considered as a syllable-timed language (cf. the psycholinguistic studies by Mehler et al.) where focus (or emphasis) accents affect initial consonant durations, and the vocalic nuclei of prosodic group last syllables tend to be lengthened (cf. the descriptive studies by Rossi, Benveniste and Mertens).

We distinguished two classes of F0 movements according to whether glissandos were present (slow overall speech rate) or not (rapid diction). Within each class, we adopted the usual distinctions between rising, atypical (*AF*) and falling (*FF*) contours. Rising continuation contours were subdivided into simple (*SRF*) and complex (*CRF*) rises; *CRF* contours are characterized by a F0 peak or resetting preceding the rise.

Analysis
The expert detected 72 prosodic boundaries (*RB*) and 23 accents (*RA*) in the six non-delexicalized extracts from the corpus. A global comparison between the distributions of the subjects' and expert's marks is given in Tab. 4.

D condition (335 marks):
The results in Tab. 5 confirm the current assumption that, in French, the lengthening of the vocalic nucleus is a reliable cue for the detection of prosodic group endings (cf. also Tab. 4).

FM and FS conditions (256 and 309 marks):
Tab. 4 and 6 suggest that glissandos help listeners to segment speech

	Nb RB	% RB	Nb RA	% RA	% UB	% UA
D	60	77	11	7	3	13
FM	57	57	5	2	8	33
FS	57	59	6	3	7	31
F+D	58	82	6	3	5	10

Nb RB/RA: Number of RB/RA detected in each condition
% RA/RB: Percentage of marks on RB/RA syllables
% UB: Percentage of marks on minor boundaries (undetected by the expert)
% UA: Percentage of marks unaccounted for by the chosen features

Table 4: Distribution of the subjects' marks in the three conditions

	VL	SL	CL
Nb syl	42	8	10
Nbm marks	4.6	3	1.3

Nb syl: Number of RB syllables marked by the subjects (per cue)
Nbm marks: Average number of marks per RB syllable (for each cue)

Table 5: Mark distribution among RB syllables - Duration cues

into prosodic groups. The high "error" rates result mainly from (cf. Tab. 7):

- confusions between *FF* and *CRF* contours,
- and erroneous delimitations of *FF* contours,

since the syllable corresponding to the end of a falling F0 contour may represent either the last syllable in the current prosodic group or the first syllable in the next group. These confusions illustrate the ambiguity of F0 contours, and the decisive contribution of rhythm to the interpretation of F0 movements in the context of speech perception. The analysis of condition **F+D** confirms this result (cf. infra).

F0+D condition (304 marks):
The joined presence of F0 and duration cues results in a significant improvement of the accuracy of boundary detection rather than in an increase of the number of detected boundaries. The confusions between *CRF* and *FF* pitch contours as well as the delimitation errors regarding *FF* contours are drastically reduced (7% of the marks vs 24% and 19% in the **FM** and **FS** conditions respectively). Moreover, the confusions between accents and boundaries are fewer in this condition than in the **D** condition (3% vs 7%). But boundaries associated with fundamental frequency (or duration) cues only are missed: 6 marked by F0 only (conditions **FM** and **FS**), 4 by duration only. This result indicates that in speech perception both F0 and duration cues are involved in the segmentation of speech into prosodic units.

C1: RB marked mainly by glissandos (23 RB)					
	FMF	SRF	CRF	AF	Nbmg
FM	0.9	2	1.8	1	1.2
FS	3	4	3.5	1.8	2.9
C2: RB marked by syllabic F0 movements (24 RB)					
FM	3.4	2.1	2.3	2	2.6
FS	2.4	1.5	0.2	1	1.5
C3: RB marked by both (19 RB)					
FM	2.8	3.5	2.4	-	2.7
FS	4.1	3	3.8	-	3.9
Total					
FM	2.3	2.4	2.1	1.5	2.2
FS	3.1	2.1	2.9	1.3	2.7

Nbmg: general average number of marks per RB and per class

Table 6: Mark distribution among RB syllables - F0 cues (average numbers of marks per RB)

	% Confusions (RB)	% Confusions (UB)	% Unaccounted
FM	16	8	9
FS	13	6	12

UB: boundaries undetected by the expert

Table 7: Percentages of errors (confusions) and unaccounted marks

6. CONCLUSION

Results show that the delexicalization procedure used with the prosodic karaoke is a useful combination for speech perception experiments, as it is more natural than many other delexicalization methods.

The subjects' performances attest that if automatic speech recognition systems could use the duration cue as well as humans, we could procrastinate for a while fundamental frequency interpretation. At last, this experiment illustrates that unlike duration, which can be interpreted individually, pitch can hardly be use alone.

7. REFERENCES

1. K.R Scherer and J.S Oshinsky. Cue utilization in emotion attributions from auditory stimuli. *Motivation and emotion*, 4:331–346, 1988.
2. F. Grosjean. Prosodic structure and word recognition. *Cognition*, 25:135–155, 1987.
3. J.R. Pijper and A.A. Sanderman. On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *J. Acoust. Soc. Am.*, 4:2037–2047, 1996.
4. T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. van der Vrecken. The mbrola project: Towards a set of high-quality speech synthesizers free of use for non-commercial purpose. In *ICSLP'96*, Philadelphia, 1996.
5. W. N. Campbell and S. D. Isard. Segment duration in a syllable frame. *Journal of Phonetics*, 19:37–47, 1991.
6. P. Langlais. *"Traitement de la prosodie dans les systèmes de reconnaissance automatique de la parole"*. PhD thesis, "Université d'Avignon et des Pays de Vaucluse", 1995.
7. V. Pagel, N. Carbonell, Y. Laprie, and J. Vaissière. Spotting prosodic boundaries in continuous speech in french. In *ICPHs'95*, volume 4, pages 308–311, Stockholm, 1995.
8. G. Fant, A. Kruckenberg, and L. Nord. Prediction of syllable duration, speech rate and tempo. In *ICSLP'92*, pages 667–670, 1992.