Perception of breaks and discourse boundaries in spontaneous speech: developping an on-line technique

Cyril Auran, Annie Colas, Cristel Portes, Monique Vion

Perception of breaks and discourse boundaries in spontaneous speech: developing an on-line technique

Cyril Auran, Annie Colas, Cristel Portes and Monique Vion

E-mail: cauran@wanadoo.fr; acolas@up.univ-mrs.fr;
cristel.portes@lpl.univ-aix.fr; mvion@up.univ-mrs.fr

Summary:

The goal of this paper is to present an on-line paradigm currently under development among the "Prosodie & discours" research group at the "Parole et Langage" laboratory, University of Provence in Aix-en-Provence. This paradigm aims at studying the perception of breaks and discourse boundaries by expert and/or naive listeners. An experimental technique was designed to compare the listeners' perception of breaks and boundaries in four different speech types (normal, low-pass filtered, re-synthezised and re-synthesized with phoneme conversion).

Thirty naive subjects took part in the present experiment, in which two versions of a one-minute tape-recorded spontaneous speech sample were compared (normal vs. converted phonemes).

The results show similar total numbers and average inter-subject agreement values for breaks identified by a majority of subjects; however, segmented portions differ both in the position of their boundaries and in the nature of the textual units they isolate.

1. Introduction

Most studies dealing with the role of prosody in the structure of spoken discourse usually rely on repeated listening and corpus annotation techniques the purpose of which is to localize prosodic breaks and boundaries and to analyse their characteristics and those of the groups of words thus set apart (Auran, Portes, Rami and Rigaud, 2001; Grosz and Hirschberg, 1992; Hirschberg and Nakatani, 1996; Lehiste, 1979; Nakatani, Hirschberg and Grosz, 1995; Portes, 2000; Portes, 2002; Strangert, 2004; Strangert and Heldner, 1995; Swerts, 1997). Studies resorting to on-line segmentation during a single listening session are far less common (Lehiste, 1979; Swerts and Geluykens, 1993; Swerts and Geluykens, 1994).

The goal of this paper is to present the first results of an on-line perception paradigm being developed among the "Prosodie & discours" research group at the "Parole et Langage" laboratory, University of Provence in Aix-en-Provence.

The project was based on a triple assumption (Vion and Colas, submitted): first, cues to the cognitive activities that occur during planning are provided by the way words are prosodically grouped; second, the resulting prosodic boundaries are used by the listener "on-line" as he/she is interpreting the produced utterances; third, the listener can use his/her spontaneous capacity in a conscious and thoughtful way to make use of prosodic cues in an experimental task.

2. Method

2.1. Task

Segmentation into prosodic segments was based on real-time perceptual analysis. Participants were instructed to press a computer key when they perceived cues for a break (or a discourse boundary). This procedure was tested on two versions of a male-speaker radio programme sample.

2.2. Speech Materials

The materials of the « original » version consists of an extract from a corpus recorded, transcribed and analysed by Cristel Portes (2004) and subsequently analysed from different perspectives by experts within the "Prosodie & discours" research group.

(1) "non je crois qu'effectiv(e)ment euh il aurait fallu pouvoir euh modifier les institutions européennes lorsque nous sommes passés de douze à quinze euh sur le fait que l'Autriche et les pays scandinaves avaient vocation à rejoindre l'union européenne personne ne pouvait l(e) mettre en en question de la même manière que pour la pologne ou la république tchèque la question qu'il fallait s(e) poser qu'on s'est posée c'est d(e) savoir si les institutions européennes étaient capables de fonctionner aussi bien à quinze vingt ou vingt-cinq qu'elles fonctionnaient à douze ça n'était pas le cas euh à l'époque euh le président de la république qui était le président mitterrand a choisi euh de fai laisser entrer l'autriche et les pays scandinaves avant de faire la réforme institutionnelle qui a été renvoyée à l'élargiss(e)ment suivant et nous nous sommes retrouvés au traité d'amsterdam qui n'est pas non plus satisfaisant et donc il faut aujourd'hui remettre l'ouvrage sur le métier"

The re-synthesized version with phoneme conversion (henceforth « saltra~dzajn ») was obtained through re-synthesis using MBROLA (Dutoit, Pagel, Pierret, Bataille and van der Vreken, 1996) and the « substitution.pl » Perl script and the « substitutions.txt » phoneme-correspondence file written by Cyril Auran (forthcoming) (see the correspondence chart for « saltra~dzajn » conversion in Table 1).

SAMPA	Replaced by
Oral vowels (i e E a A O o u y 2 9 @)	a
Nasal vowels (e~a~o~9)	a~
Semi-consonants (j w H)	j
Voiceless stops (p t k)	t
Voiced stops (b d g)	d
Voiceless fricatives (f s S)	s
Voiced fricatives (v z Z)	z
Liquid (l)	l
Liquid (R)	R
Nasals (m n N)	n

Table 1: Correspondence chart for « saltra~dzajn » phoneme conversion.

2.3. Procedure

The experiment was run on an Apple iBook G4 computer under Mac os X (10.3), and took place in an anechoic room. The participants were provided with headphones and did the task individually. Sound file listening and on-line annotation rely on software by Cyril Auran (forthcoming) (tcl/tk programming language exploiting Wish). The system evaluates the subject's base reaction time through a series of 10 (pre-recorded) irregular beeps. Script « testeur.tcl » more particularly allows the subject to segment the speech signal on-line: the position (P) of each response is recorded with reference to the beginning of the sound file, a corrected value (cP) taking the subject's base reaction time into account is computed

and a TextGrid file (cf PRAAT: Boersma, & Weenink, 2005) summarizing both those values is eventually generated.

The experimenter began by exposing the purpose of the study (study of the prosodic cues used by listeners in order to organize and understand spoken language) and presenting the task to each subject. The task itself consists in the subjects listening to a recorded sequence of speech and, basing their decisions on the "music of speech", segmenting this sequence on-line into blocks. A block, was defined as a sequence of speech, i.e. a word or a group of words which seem to "hold together" (Strangert, 2004). The experiment then continues in two stages. In the first stage (off-line), the subject is made to listen to 4 short sequences through headphones. For each sequence, the subject is required to tell how many blocks they heard. During stage two (on-line), the subject is made to familiarize with the response procedure which consists in signalling block boundaries on-line by pressing the spacebar on the computer's keyboard. The experimenter particularly insists on the fact that the subject's reaction time be very short in order not to interfere with upcoming speech. The experiment is then run in 3 parts: (1) measurement of the subject's base reaction time; (2) familiarization to the speaker's voice and segmentation training on two (original condition) or three (saltra~dzajn condition) 10-second speech sequences; (3) segmentation of a single 1-minute speech sequence.

The experiment lasted approximately 15 minutes, at the end of which the participants were asked to detail what criteria they had used to isolate a group of word as a unit.

2.4. Participants

The participants were non-linguist native French speakers. They were young adult students at the University of Provence in Aix-en-Provence, France. The perceptual evaluation of the materials required a total of 30 participants (15 participants x 2 conditions).

2.5. Data Collection Design

Each participant listened to, and had to segment only one of the two versions.

2.6. Predictions

First, given the fact that the task implied on-line treatment during a single listening session whereas the experts could resort to repeated listening of the original recording, the experimental segmentation is not expected to be as fine-grained as the experts'.

Second, taking into account the fact that, in the original condition, participants could base their decisions on both prosodic and syntactic-semantic information, whereas in the saltra~dzajn condition, they only had prosodic information at their disposal, greater precision in the segmentation is to be expected in the former condition.

3. Results

3.1. Database

A single grid was designed for each version grouping the auditory evaluations of all 15 listeners. Thus the grid displayed the whole information about the location of the breaks and boundaries they noted.

In this perspective, each recording was pre-segmented using Praat (Boersma & Weenink 2005): 46 intervals were thus isolated, following the original segmentation by Portes (2004). For each version, these intervals were used as a reference grid for the analysis of the experimental data. For each participant, the position of each response (cP) was associated with the relevant reference interval ; a mean response position was then computed for each interval. When, for a single participant, several cPs were associated with a single interval, those cPs were ordered within the interval according to their values and several mean response positions were computed for this interval. The whole set of mean response positions was then used to determine the positions of boundaries within the global TextGrid file associated with each group.

3.2. Overall results

Table 2, given the number of participants involved, summarizes the main characteristics of the segmentation obtained for each version.

version	original	saltra~dzajn
Total number of intervals	**49**	**54**
Mean number of boundaries	**20.27**	**20**
Total number of word-internal boundaries	**11**	**28**

Table 2: All boundaries

A total of 54 intervals was identified for the saltra~dzajn version, and 49 for the original version. Taking into account the fact that, 6 intervals in the saltra~dzajn version correspond to a silent stretch, both versions contain a similar number of textual segments. The mean total number of boundaries for the saltra~dzajn version is 20, and 20.27 for the original version; once again, these results can be regarded as similar for both versions.

However, the two versions differ regarding the position of boundaries: indeed, for the saltra~dzajn version, 28 boundaries (i.e. 51.9% of all boundaries for this version) are word-internal whereas such a phenomenon is less frequent for the original (18 boundaries, i.e. 36.7% of all boundaries for this version).

The proportion of listeners agreeing on the location of a given break was computed, and only those breaks identified by at least 8 listeners out of 15 (majority) were kept (henceforth « majority boundaries »). Table 3 presents the specific characteristics of these majority boundaries for both versions.

version	original	saltra~dzajn
Total number of majority boundaries	**15**	**15**
Mean inter-speaker agreement for majority boundaries	**10.1**	**10.6**

Table 3: Majority boundaries

The total number of majority boundaries is identical for both versions (15 boundaries). Mean inter-speaker agreement on these boundaries is also similar for both versions (saltra~dzajn version = 10.6, original version = 10.1).

3.3. Comparison of the segmentations

Table 4 presents the segmentation of each text taking into account the position of majority boundaries. Each time a word-internal boundary is met, the relevant word appears in capital letters. Table 4 also presents the results of the off-line analysis by 3 experts from the "Prosodie & discours" research group. Each expert had to rate boundary strength using a four-point scale, with "1" meaning 'very weak boundary" and "4" "very strong boundary". This scale was inspired from the ToBI standard for prosodic transcription ("Tone and Break Indices": (Silverman, Beckman, Pitrelli, Ostendorf, Wightman, Price, Pierrehumbert and Hirschberg, 1992). 17 boundaries were rated either 3 or 4 by all experts. Words preceding these consensual boundaries appear in bold type.

original version	« saltra~dzajn » version
non je crois **qu'effectiv(e)ment**	non je crois **qu'effectiv(e)ment**
euh	euh Il aurait fallu POUV
il aurait fallu pouvoir euh	VOIR euh modifier les institutions européennes lorsque nous sommes passés de douze à **quinze**
modifier les institutions européennes lorsque nous sommes passés de douze à **quinze**	**euh** sur le
euh sur le fait **que**	fait **que** L'AU
l'autriche et les pays **scandinaves**	TRICHE et les pays **scandinaves**
avaient vocation à rejoindre l'union **européenne**	avaient vocation à rejoindre l'union **européenne** personne ne pouvait l(e) mettre en en question de la même manière que pour la pologne ou la rèpublique **tchèque**
personne ne pouvait l(e) mettre en en question dla même manière que pour la Pologne ou la République **tchèque**	la question qu'il fallait s(e) poser qu'on s'est **posée** c'est d(e) savoir si les institutions **europèennes** É
la question qu'il fallait s(e) poser qu'on s'est **posée** c'est d(e) savoir si les institutions **européennes** étaient **CAP**	Etaient **capables** de fonctionner aussi bien à quinze vingt ou vingt-cinq qu'elles fonctionnaient à douze ça n'était pas le **cas**
PABLES de fonctionner aussi bien à quinze vingt ou vingt-cinq qu'elles fonctionnaient à douze ça n'était pas le **cas**	euh à l'époque euh le président de la rèpublique qui était le président **mitterrand** a **CH**
euh à l'époque euh le	**CHoisi** euh
président de la république qui était Le président **MITTERRAN**	euh de fai laisser entrer l'autriche et les pays scandinaves avant
AND a **choisi** euh de fai laisser entrer l'autriche et les pays SCANDINAV	de faire la réforme institutionnelle qui a É
AVES avant de faire la réforme institutionnelle qui a été renvoyée à l'élargiss(e)ment **suivant** et nous nous sommes retrouvés au traité **d'Amsterdam** qui n'est pas non plus **satisfaisant**	TÉ renvoyée à l'élargiss(e)ment **suivant** et nous nous sommes retrouvés au traité **d'amsterdam** qui n'est pas non plus **satisfaisant** et DON
et donc il faut aujourd'hui remettre l'ouvrage sur le métier	ONC il faut aujourd'hui remettre l'ouvrage sur le métier

Table 4: Text segmentation for each version (majority boundaries)

At first sight, the segmentations, which are very similar regarding their total numbers of intervals, greatly differ with respect to the position of boundaries (only 6 boundaries in common – excluding the end of the text) and to the segments isolated by the participants.

Moreover, as already mentioned for raw data, the two versions differ with respect to the precision of boundary placement: 6 word-internal boundaries in the saltra~dzajn version, against only in the original version. Close examination of these words show that, in the the original version, one word is split on the boundary between two syllables (ca/pable) and two words on the vowel in the last syllable: in these last two cases, the word is phrase-terminal (/a/ and /a~/ in « les pays scandinaves » and « le président mitterand » respectively). In the saltra~dzajn version, two words are split on the boundary between 2 syllables (au/triche and é/té) and four within a syllable, twice on the onset consonant (pouv/voir and ch/choisi) and twice on the vowel (don/onc and é/taient).

Examination of the the textual segments related to the experts' off-line segmentation (bold-face words in table 4), shows a 9-boundary match in the original version and a 5-boundary match in the saltra~dzajn version. Moreover, nearly all cases of word-internal boundaries can be explained with referece to the experts' boundaries: in the normal version, 2 word-internal boundaries out of 3 are located one syllable before a high-rated (strong) boundary identified by the experts; in the saltra~dzajn version, 5 word-internal boundaries out of 6 are located one syllable after a high-rated (strong) boundary identified by the experts.

To sum up, the segmentation of the original version coincides with the experts' for 64% of its boundaries and displays 2 cases of anticipated segmentation (in total: 78.6 % agreement with the experts). The segmentation of the saltra~dzajn version coincides with the experts' for 35. % of its boundaries and displays 5 cases of delayed segmentation (in total: 71.4 % agreement with the experts).

4. Conclusion

As expected, the participants provided a less detailed segmentation than that by the experts, but one quite congruent with it since the two on-line segmentations show an average agreement of 75% with the boundaries identified by the experts. In line with our predictions, the segmentation of the original version, for which listeners had access to both prosodic and syntactic-semantic types of data, is closer to the experts' and sometimes anticipates the presence of a boundary identified by them; conversely, the segmentation of the saltra~dzajn version, where listeners could only access prosodic information usually displays a slight delay in the location of boundaries with reference to those set by the experts.

The reliability of the segmentation and the presence of anticipated and delayed responses depending on the version used remain to be explored in relation with the analysis of the segmentations of the low-pass filtered and resynthesized versions of the original recording.

Such preliminary results seem to constitute rather encouraging elements in favour of the development and systematisation of a real-time experimental approach to the prosodic cues of discourse structure.

References

Auran, C., Portes, C., Rami, E. and Rigaud, N.: 2001. 'La distinction entre frontières conclusives et continuatives est-elle pertinente dans le discours spontané?' in, *Journées Prosodie 2001*, Grenoble.

Boersma, P. and Weenink, D.: 2005, 'Praat. Doing phonetics by computer [computer program]. Available from http://www.praat.org.'

Dutoit, T., Pagel, V., Pierret, N., Bataille, F. and van der Vreken, O.: 1996. 'The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes', in, *ICSLP'96*, Vol. 3, 3 vols., Pliladelphia, pp. 1393-1396.

Grosz, B. and Hirschberg, J.: 1992. 'Some intonational characteristics of discourse structure.' in, *Proceedings International Conference on Spoken Language Processing (Banff)*, Vol. 1, pp. 429-432.

Hirschberg, J. and Nakatani, C. H.: 1996. 'A prosodic analysis of discourse segments in direction-giving monologues.' in, *Thirty-fourth Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, pp. 286-293.

Lehiste, I.: 1979. 'Perception of sentence and paragraph boundaries', in B. Lindblom and S. Ohman, (eds.), *Frontiers of speech communication research*, Academic Press., New York:, pp. 191-201.

Nakatani, C. H., Hirschberg, J. and Grosz, B. J.: 1995. 'Discourse structure in Spoken Language: Studies on Speech Corpora', in, *AAAI-95 Spring Symposium o empirical Methods in Discourse Interpretation*, Palo Alto, C.A., pp. 106-112.

Portes, C.: 2000, *Approche du rôle de la prosodie dans la structuration du discours oral en Français*. research note. Université de Provence.

Portes, C.: 2002, 'Approche instrumentale et cognitive de la prosodie du discours en français.' *Travaux Interdisciplinaires Parole et Langage* **21**, 101-119.

Portes, C.: 2004. 'Prosodie et économie du discours : spécificité phonétique, écologie discursive et portée pragmatique de l'intonation d'implication.' in, Université de Provence.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C. M., Price, P., Pierrehumbert, J. and Hirschberg, J.: 1992. 'ToBI: a standard for labeling English prosody', in, *second Iinternational Conference on Spoken Language Processing*, Vol. 2, pp. 867-870.

Strangert, E.: 2004. 'speech chunks in conversation:syntactic and prosodic aspects', in, *Speech Prosody*, pp. 305-308.

Strangert, E. and Heldner, M.: 1995, 'Labelling of boundaries and prominences by phonetically experienced and non-experienced transcribers.' *PHONUM* **3**, 85-109.

Swerts, M.: 1997, 'Prosodic features at discourse boundaries of different strength.' *Journal of the Acoustical Society of America* **101**, 514-521.

Swerts, M. and Geluykens, R.: 1993, 'The prosody of information units in spontaneous monologue.' *Phonetica* **50**, 189-196.

Swerts, M. and Geluykens, R.: 1994, ' Prosody as a marker of information flow in spoken discourse.' *Language and Speech* **37**, 21-43.

Vion, M. and Colas, A.: in press, La planification des unités prosodiques dans la narration : contrôle intentionnel et contraintes opértionnelles.' *Travaux Interdisciplinaires Parole et Langage* **24**.