Speech Prosody 2016
31 May - 3 June 2016, Boston, USA

Perception of Prosodic Boundaries by Naïve Listeners in French

Anne Catherine Simon[1], *George Christodoulides*[1]

[1]Université catholique de Louvain, Belgium

anne-catherine.simon@uclouvain.be, george@mycontent.gr

Abstract

We present the results of an experiment on the on-line perception of prosodic boundaries by 84 naïve listeners. Potential samples from a multi-genre corpus of spoken French were stratified based on 3 prosodic measures, and 48 samples (mean length 29.9 seconds) were selected, balanced for their degree of fluency. Each sample was resynthesized to obliterate lexical content while keeping its syllabic structure and intonation. Four sets of stimuli were created (12 natural, 12 manipulated speech). Each sample was presented only once to 20 to 22 participants, who were instructed to press the space-bar as soon as they heard the end of a "group of words". Baseline reaction time to simple tones was measured before and after the perception task. In total, 17195 perceived prosodic boundaries (PPB) were recorded. For each PPB, we calculated its strength, the temporal delay and dispersion of responses. Results show that although the number of PPBs is similar in natural speech (NS) and manipulated speech (MS), the types of PPBs, their acoustic correlates and relation to syntax vary between the two conditions; in NS, we show that the presence of a filled pause and the syntactic structure act as strong cues to PPBs.

Index Terms: prosodic boundaries, on-line perception experiment, prosody-syntax interface, disfluencies

1. Introduction

Prosody is known to be central to language comprehension by helping listeners segment the incoming text (for example [9], [30], [14]). However, there is no consensus on a segmentation method that could be applied, either manually or automatically, to large corpora of speech and the factors contributing to the perception of prosodic boundaries are still investigated. Our research, therefore, has the following four objectives: first, examine the degree of consensus in the perception of prosodic boundaries by non-expert (naïve) listeners, with a view to modelling this behaviour, and using such models for automatic annotation. Second, study the variation in the perception of prosodic boundaries across different conditions (speaking style, natural vs. manipulated speech, individual differences). Third, compare the perceived prosodic boundaries (PPBs), as identified by naïve listeners under realistic listening conditions, with the PPBs annotated by experts. And finally, study the role of different acoustic and syntactic cues in the perception of prosodic segmentation.

2. Related Work

Many corpora of spoken language developed in the past 50 years often include some sort of prosodic segmentation into "intonation units" or annotation of "prosodic boundaries" (for example: "tone units" in the London-Lund Corpus of Spoken English [27]; "intonation units" in the Santa Barbara Corpus of American English [13] or in the Aix-Marsec Corpus of Spoken British English [2]; "*périodes intonatives*" in the Rhapsodie corpus of Spoken French [23], "major intonation units" in the LOCAS corpus [11]). Those prosodic segments are used to explore the prosody-syntax-discourse interface; they are either manually annotated by experts, or automatically detected based on acoustic features. We would like to take a step back and test how these "prosodic units" are perceived on-line by listeners engaging in language comprehension and discourse processing, as opposed to experts who listen carefully to an excerpt in order to annotate it. We focus on the variability in the perception of prosodic boundaries, in line with recent research that emphasises individual differences in language comprehension. In previous work, two techniques have been used to tackle the fact that not all participants perceive boundaries at the same time/locations. In some studies, consensus (or majority) boundaries are identified at those locations where a certain proportion of the participants have identified a boundary (e.g. 67% in [25], 8 or more out of 12 subjects in [3]). Alternatively, the "boundary strength" is defined as the proportion of subjects indicating that they perceived a PB at a given location, expressed as a value between 0 and 1 (e.g. [28]:516; [8]:1152). We have chosen the latter method for the analysis presented here.

2.1. Acoustic cues to prosodic boundaries

We reviewed previous experimental studies on perceived prosodic boundaries (PPBs) in different languages (Dutch, English, French, Hebrew and Swedish) in order to identify the acoustic cues to PBs, the impact of syntax on the perception of PBs, and the role of disfluencies on the perception of PBs. In general, the acoustic features of units bearing a PPB are compared to those of other units where no PB had been perceived (e.g. the final syllable of each word in the sample in [8], the end of intonation units in [25], or prosodic phrases in [28], as predicted by a phonological model or annotated by an expert). Because of those methodological differences it difficult to compare previous studies; however the following trends emerge.

In English map-task and broadcast extracts, Smith [25] showed that pauses (a broad category including silent and filled pauses, as well as breathing, longer than 150 ms) favour the perception of a boundary, but are not a reliable indicator. Studying Dutch spontaneous monologs (picture descriptions) Swerts [28] observes a trend for longer pauses (silent pauses longer than 250 ms) to be associated with stronger perceived boundaries. Silent pauses appear as a major device used to mark boundaries, while filled pauses are more ambiguous cues to this respect (see section 2.3).

Vowel duration, as well as intensity, are often correlated to prominence and stressed syllables. It is therefore much more

10.21437/SpeechProsody.2016-238

language dependent, as stress and boundary tones are located on the same syllable in stress-group languages like French, but not in lexical-stress languages like English. In conversational English, measurements indicate that vowel duration is a robust correlate of perceived phrase boundary: most stressed vowels in pre-boundary locations are significantly longer than those that are not, according to Mo [22]. In other studies however, no significant difference has been found between the mean duration of (segments in the) words preceding perceived boundaries and other words, and sometimes the mean duration of words was significantly shorter (in English, [25]). In Swedish conversational speech, Strangert [26] reports that the durations of words and word-final rhymes are generally longer before weak boundaries than before strong, and that the length decreases with the number of words in the chunk.

Studies examining the role of f_0 contour generally show that there is no strong or stable correlation between low vs. high pitch and the perception of boundaries, though in French, Portes [24] observes that weak PB are associated with rising contours and strong PB with falling contours. De Pijper [10] found that melodic discontinuity was the only phonetic cue to systematically occur in isolation for marking PB, which is consistent, while Smith [25] shows that the magnitude of f_0 movement better correlates with PB strength than the direction of the contour.

2.2. The role of disfluencies

It has been shown that disfluencies are not always consciously processed: perceptual experiments have shown how hearers systematically displace within-constituent hesitations to constituent boundaries [18]. The influence of filled pauses on the perception of prosodic boundaries is a source of substantial individual variation, leading some authors to discard those cases from their analysis, because such PPBs are less consensual [1]. Other studies show that final word lengthening due to hesitation was taken into account by listeners in their decision on boundary perception, and impeded them to perceive a boundary [25].

2.3. The influence of syntax and semantics in the perception of prosodic boundaries

When groups of subjects are asked to annotate discourse units (paragraphs, sentences, etc.) on the basis of only a transcript, it has been shown that they are less consistent (i.e. produce less consensus boundaries) than groups of subjects that also have access to the speech signal [28], [16]. Subjects working on the transcript alone also annotate fewer boundaries than those both reading the transcript and listening to the speech, except at the higher level of the discourse hierarchy [29].

On the other hand, subjects labelling prosodic boundaries in speech in their own language perceive more PBs than subject annotating delexicalised speech or speech in a language they don't understand [17], [3], [21]. This seems indicate that prosodic cues can help disambiguate among alternative segmentations of the same text, and listeners combine prosodic and syntactic cues to segment discourse.

3. Method

3.1. Perceptual Experiment Design and Hypotheses

We designed a perceptual experiment in which participants were listening to a short sample of speech and were instructed to press a key whenever they perceived the end of a "group of words" (this instruction was deliberately vague, in order to avoid biasing subjects towards a syntax-based analysis). Participants could only listen to each sample once and the collection of responses was done in real time, in order to be as close as possible to natural conditions of speech perception and comprehension. On the other hand, using such an experimental protocol, it is not possible to ask participants to indicate the perceived "strength" of each PB; additionally individual differences in motor and co-ordination skills should be taken into account. Participants were asked to annotate 12 stimuli of natural speech and subsequently 12 stimuli of manipulated speech (the delexicalisation process is described below). Our hypotheses were the following: (1) PB perception does not rely on prosodic cues only but is also influenced by syntax and semantic features; (2) fewer PBs will be detected under in the manipulated speech stimuli compared to the natural speech stimuli; (3) under the natural speech condition, more PBs are perceived at the end of syntactic units; (4) untrained non-expert listeners under naturalistic conditions will detect fewer PBs than trained expert (only verifiable under the NS condition); (5) pauses are the strongest cue to PB perception.

3.2. Stimuli Preparation

3.2.1. Selection

The speech stimuli were extracted from the LOCAS-F corpus [11]. A database of potential stimuli was prepared, containing monological inter-pausal units, 20-60 seconds long. In order to reach a balanced set of stimuli with respect to their acoustic parameters, those potential stimuli were clustered according to 4 criteria (articulation rate, silent pause ratio, melodicity, filled pauses to number of syllables ratio), using k-means clustering. We created a stratified selection of stimuli in two groups (fluent vs. disfluent); 4 groups of 12 stimuli were selected, with an average duration of 29.9 seconds (min: 5.1, max 39.9).

3.2.2. Manipulation

We produced corresponding manipulated speech stimuli, in order to mask lexical content, while retaining the temporal, syllabic and intonation structure. Phonemes were randomly replaced with another phoneme from the same group (plosives, fricatives, nasals, liquids, glides, vowels, nasal vowels), ensuring that resulting diphones exist in French. Phone duration was kept intact, while the intonation contour was approximated (10 points). The manipulated stimuli were then synthesised using the MBROLA TTS system. The resulting stimuli sound similar to a pseudo-language (compared to the hum resulting from band-pass filtering methods).

3.3. Procedure

In total, 88 university students took part in the perceptual experiment. They were all studying at the faculties of Psychology and Modern Languages at the University of Louvain in Belgium, and had no previous experience in prosodic annotation. The experiment was conducted in the computer labs of the faculty, and lasted approximately 30 minutes. No participant reported a hearing problem, but 4 were excluded from the final analysis (2 were non-native speakers of French and 2 did not finish the experiment).

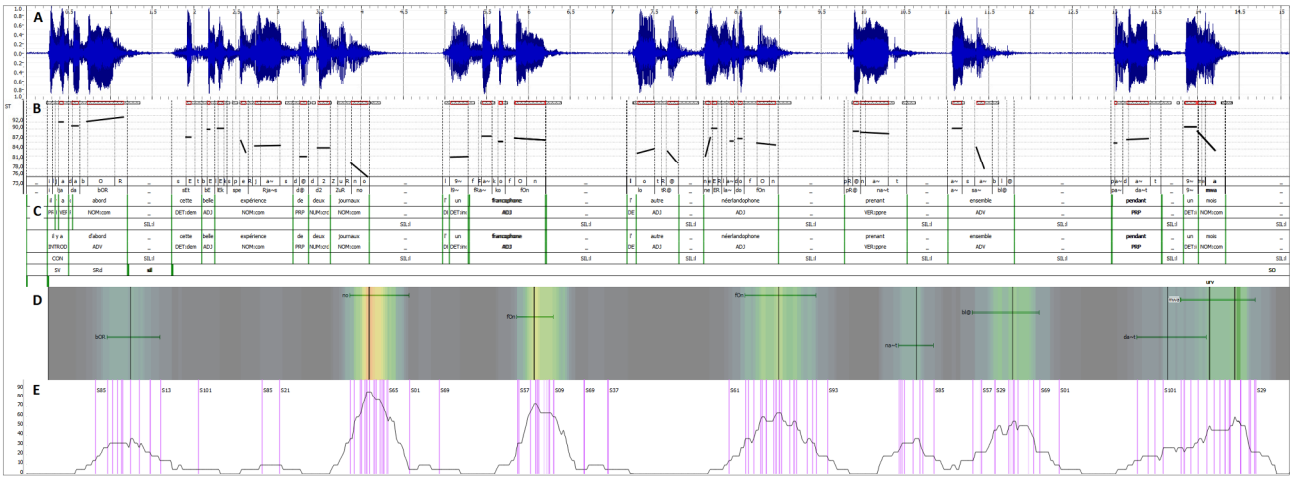

Figure 1. Visualisation of experimental results in *Praaline*. The waveform (A) is displayed along with its *Prosogram* (B) and transcription, POS and syntactic annotation (C). Subject responses (centred) are shown in panel E, along with a moving average of the number of responses. Local maxima are selected as the PPB locations; panel D displays the extent of each group of responses considered as part of the same PPB, in order to calculate the force, dispersion and mean delay.

The experimental sequence ran as follows: participant identification, working memory capacity test, tonal acuity test, baseline response time test (participants were asked to press the key as soon as they heard a pure tone); training; segmentation of natural stimuli; segmentation of manipulated stimuli; repetition of the baseline response time test. The experiment was presented using *OpenSesame* [19].

3.4. Data Analysis

The main procedure for analysing the raw data is visualised in Figure 1. For each subject, we calculated a mean RT from their responses to the pure tones. These values were subtracted from their responses in order to centre them with respect to a potential location of a PB and to reduce variability induced by individual motor skill differences. A moving average (window size 250 ms) of the number of responses was calculated and the local maxima of this value were considered as the PPB sites. In order to group subject responses correlated with a PPB, we followed the following algorithm: starting from the centre and within a window of 500 ms on either side, a response is attributed to the PPB if its distance from the previous response is less than 300 ms; we are thus attempting to detect clusters of responses triggered by the same cues. These responses are subsequently treated as a group: each group gives rise to a PPB. These PPBs were correlated with the nearest final syllable of a token (PPB sites falling within a silent pause were attributed to the previous final syllable). For each PPB we calculate three measures: the boundary force is the proportion (%) of participants who registered a response at this PPB site; the boundary delay is the arithmetic mean of the temporal difference between the syllabic nucleus and each subject response; and the boundary dispersion is the standard deviation of the aforementioned temporal differences (response times). Furthermore, the corpus contains detailed annotations: part-of-speech tags (using *DisMo* [6]), a manual syntactical annotation in functional sequences and dependency clauses, acoustic/prosodic features extracted using *Prosogram* [20] for each syllable, as well as a perceptual disfluency annotation for the stimuli. We used *Praaline* [5] to process, visualise and manage this multi-level annotation (Figure 1).

4. Results

In total 17195 responses were registered, grouped into 1270 perceived prosodic boundaries. Subjects perceived on average 8.75 prosodic boundaries per stimuli in natural speech (NS) and 8.6 in manipulated speech (MS). A comparison of the boundary force of PPBs with the corresponding expert annotation (on the same PB locations) can be seen in Figure 2.

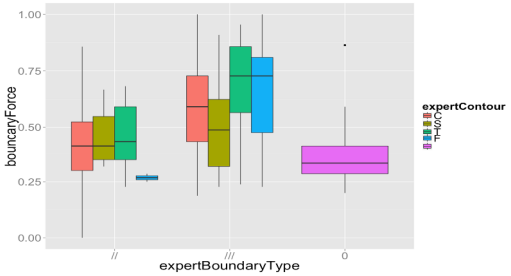

Figure 2. *Perceived boundary force, compared to the expert annotation*

Expert annotators labelled the corpus for two boundary strengths: weak (//) corresponding to 93 PPBs in NS, and strong (///) corresponding to 451 PPBs in NS, while 85 PPBs in NS were attributed to syllables the experts did not annotate as boundaries (0). Contours were annotated by the experts as follows: C for rising; S for level; T for falling; F for rising-falling. We observe that the mean perceived force of PPBs annotated as strong by the experts was significantly higher than that of PPBs annotated as weak (Cohen's $\delta = 0.92$); that falling-contour PPBs had the highest mean perceived force; and that level-contour PPBs have a similar distribution of perceived force, regardless of whether the experts annotated them as weak or strong. Furthermore, Figure 3 shows the relationship between the attribution of a PPB to a token and its POS tag: we confirm that more and stronger boundaries are perceived primarily on lexical items.

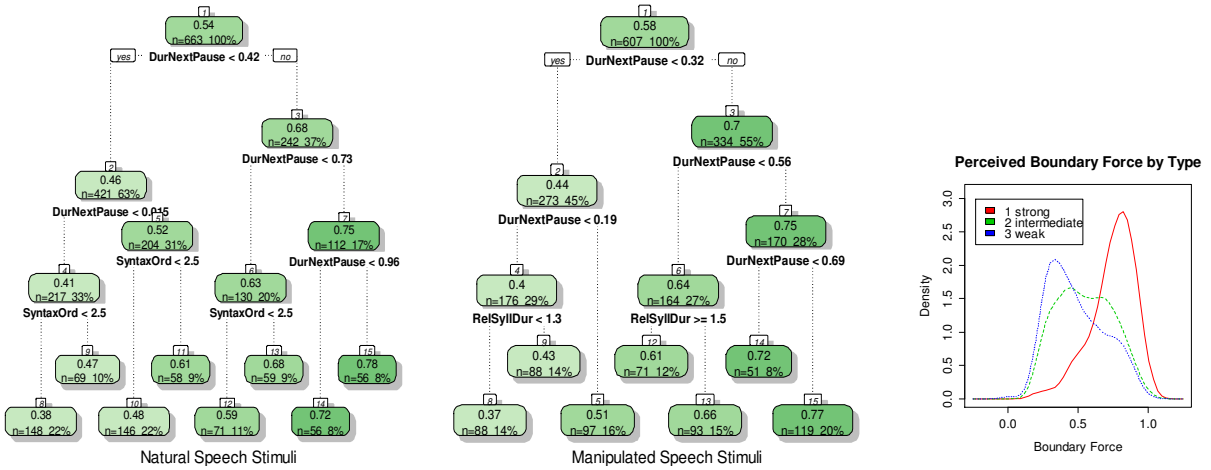

Figure 4: *Regression decision trees of Boundary Force for natural speech stimuli (left) and manipulated speech stimuli (middle). After applying k-means clustering to the acoustic features of the PPBs, three groups arise, with corresponding boundary force distributions (right).*

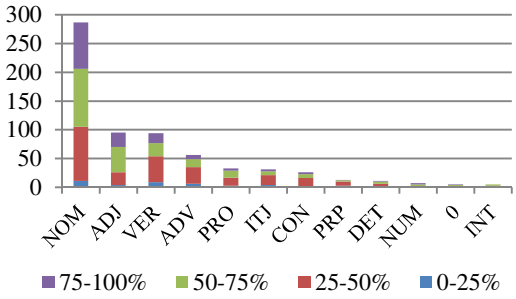

Figure 3. *Boundary force / POS tag*

In order to study the relative importance of acoustic and syntactic cues, we performed three statistical analyses. A regression tree was fitted, using subsequent pause duration, relative syllable duration, relative pitch and syntactic boundary strength as the predictors of boundary force; as can be seen in Figure 4 (left, middle), pauses are the strongest cue of PBs, followed by syntax which (as expected) only influenced the natural speech condition. We also applied k-means clustering to the aforementioned acoustic features of the PPBs. The optimal number of clusters resulted to be 3. When PPBs were assigned their cluster, and the distribution of the boundary force is plotted for each cluster (Figure 4, right) we notice that the clusters naturally represent three types of boundaries: weak, intermediate and strong ones. Finally, we tested linear regression models, with the boundary force as the dependent variable and the acoustic / syntactic cues as the predictors. The results can be found in the following two tables.

Table 1. *Linear regression model coefficients for natural speech stimuli boundary force*

Coefficients	Estimate	Pr(>|t|)
(Intercept)	0.3217475	< 2e-16 ***
DurNextPause	0.2509497	< 2e-16 ***
RelSyllDur	0.0596647	0.00127 **
RelPitch	-0.0008602	0.70598
Syntax = 0	-0.0170935	0.75661
Syntax = MD	0.0642374	0.08470
Syntax = REC	0.1029944	4.83e-11 ***
Syntax = SEQ	0.0409998	0.02971 *

Table 2. *Linear regression model coefficients for manipulated speech stimuli boundary force*

Coefficients	Estimate	Pr(>|t|)
(Intercept)	0.385792	<2e-16 ***
DurNextPause	0.292958	<2e-16 ***
RelSyllDur	0.039630	0.0411 *
RelPitch	0.001308	0.5795
Trajectory	0.004466	0.0356 *

In both cases, silent pauses are the most important cue to PB perception. In the natural speech condition, clause boundaries (Syntax=REC) follow immediately. Relative syllable duration (i.e. lengthening) is also significant, while we fail to reach significance for relative pitch (pitch movements). These results have to be interpreted in light of the prosodic structure of the French language (cf. section 2). A more detailed description of the acoustic and syntactic features used, can be found in our previous study [7].

5. Conclusions

We have presented an experiment to study the online perception of prosodic boundaries by non-expert listeners. We have confirmed the following of our hypotheses: (1) PB perception does not rely on prosodic cues only but is also influenced by syntax and semantic features; (2) fewer PBs will be detected under in the manipulated speech stimuli compared to the natural speech stimuli; (3) under the natural speech condition, more PBs are perceived at the end of major syntactic clauses, while smaller syntactical units are less important cues for segmentation; (4) untrained non-expert listeners under naturalistic conditions will detect fewer PBs than trained expert and (5) pauses are the strongest cue to PB perception. Further analysis of the data, not presented here due to lack of space, will focus on the role of disfluencies and individual differences, and in the correlation of these individual differences with the measures of working memory capacity and pitch perception. We also plan to re-run the experiment with a larger pool of participants.

6. References

[1] Amir N., Silber-Varod V., Izre'el S., "Characteristics of intonation unit boundaries in spontaneous spoken Hebrew – perception and acoustic correlates" In *Proceedings of Speech Prosody 2004*, Nara, Japan, 2004, pp. 677–680.

[2] Auran C., Bouzon C., Hirst D., "The AixMARSEC project: an evolutive database of spoken English," in In Bel, B. & Marlien, I. (eds) *Proceedings of the Second International Conference on Speech Prosody*, 2004, pp. 561–564.

[3] Auran C., Colas A., Portes C., Vion M., "Perception of breaks and discourse boundaries in spontaneous speech: developing an on-line technique," in *Proceedings IDP 2005*, Aix-en-Provence, 2005, 7 p.

[4] Campbell N., "Automatic detection of prosodic boundaries in speech," *Speech Commun.*, vol. 13, no. 3–4, pp. 343–354, 1993.

[5] Christodoulides G., "Praaline: Integrating tools for speech corpus research", *Proc. of the 9th International Conference on Language Resources and Evaluation (LREC)*, pp. 31-34, 2014.

[6] Christodoulides G., Avanzi M., Goldman J.-Ph., "DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator", *Proc. of 9th International Conference on Language Resources and Evaluation (LREC)*, pp. 3902-3907, 2014.

[7] Christodoulides G., Simon A.C., "Exploring Acoustic and Syntactic Cues to Prosodic Boundaries in French A Multi-Genre Corpus Study", *Proceedings of the 18th International Congress of Phonetic Sciences*, 10-14 August 2015.

[8] Cole J., Mo Y., Baek S., "The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech," *Language and Cognitive Processes*, vol. 25, no. 7–9, pp. 1141–1177, 2010.

[9] Cutler A., Dahan D., van Donselaar W., "Prosody in the Comprehension of Spoken Language: A Literature Review," *Language and Speech*, vol. 40, no. 2, pp. 141–201, Jan. 1997.

[10] de Pijper J. R., Sanderman A. A., "On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues," *The Journal of the Acoustical Society of America*, vol. 96, no. 4, pp. 2037–2047, 1994.

[11] Degand, L., Martin, L.J., Simon, A.C. 2014. Unités discursives de base et leur périphérie gauche dans LOCAS-F, un corpus oral multigenres annoté. *Actes du 4ème Congrès Mondial de Linguistique Française* 2014, Berlin, Germany: EDP SciencesC.

[12] DeLooze, C., Hirst D., "Detecting changes in key and range for the automatic modelling and coding of intonation," in In Speech Prosody 2008.

[13] DuBois J. W., Schuetze-Coburn S., Cumming S., Paolino D., "Outline of discourse transcription," in *Talking data: Transcription and coding in discourse research*, J. A. Edwards and M. D. Lampert, Eds. Hillsdale, N.J.: Erlbaum, 1993, pp. 45–89.

[14] Frazier L., Carlson K., Clifton J., "Prosodic phrasing is central to language comprehension," *Trends in Cognitive Sciences*, vol. 10, no. 6, pp. 244–249, Jun. 2006.

[15] Goldman J.-P., Auchlin A., Simon A. C., "Discrimination de styles de parole par analyse prosodique semi-automatique," in *Actes d'IDP 2009*, Paris, 9-11 septembre 2009, Paris, 2011, pp. 287–301.

[16] Hirschberg J., Nakatani C. H., Grosz B. J., "Conveying Discourse Structure through Intonation Variation," in *SDS-1995*, Denmark, 1995, pp. 189–192.

[17] Kreiman J., "Perception of sentence and paragraph boundaries in natural conversation," *Journal of Phonetics*, no. 10, pp. 163–175, 1982.

[18] Martin J. G., Strange W., "The perception of hesitation in spontaneous speech," *Perception & Psychophysics*, vol. 3, no. 6, pp. 427–438, 1968.

[19] Mathôt S., Schreij D., Theeuwes J. "OpenSesame: An open-source, graphical experiment builder for the social sciences,", *Behavior Research Methods*, vol. 44, no. 2, pp. 314-324, 2012.

[20] Mertens P., "The ProsoGram: Semi-automatic transcription of prosody based on a tonal perception model". In *Proceedings of Speech Prosody 2004*, Nara, Japan, 2004.

[21] Mettouchi A., Lacheret-Dujour A., Silber-Varod V., 'el Shlomo I., "Only prosody? Perception of speech segmentation in Kabyle and Hebrew," *Nouveaux Cahiers de Linguistique Française*, no. 28, pp. 207–218, 2007.

[22] Mo Y., "Duration and intensity as perceptual cues for naïve listeners' prominence and boundary perception," in *Proceedings of the 4th Speech Prosody Conference*, Campinas, Brazil, 2008, pp. 739–742.

[23] Pietrandrea P., Kahane S., Lacheret-Dujour A., Sabio F., "The notion of sentence and other discourse units in corpus annotation," in *Spoken Corpora and Linguistic Studies*, T. Raso and H. Mello, Eds. Amsterdam / Philadelphia: John Benjamins, 2014, pp. 331–364.

[24] Portes, C., "Approche instrumentale et cognitive de la prosodie du discours en français", *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA)*, Laboratoire Parole et Langage, no. 21, pp.101-119, 2002.

[25] Smith C. L., "Naïve listeners' perceptions of French prosody compared to the predictions of theoretical models," in *Actes d'IDP 2009, Paris, 9-11 September 2009*, Paris, 2011, pp. 355–349.

[26] Strangert E., "Speech Chunks in Conversation: Syntactic and Prosodic Aspects," in *Proceedings of Speech Prosody* 2004, Nara, Japan, 2004, pp. 305–308.

[27] Svartvik J., Quirk R., *A Corpus of English Conversation*. Lund: Lund Univeristy Press, 1980.

[28] Swerts M., "Prosodic features at discourse boundaries of different strength," *Journal of the Acoustical Society of America*, no. 101, pp. 514–521, 1997.

[29] van Donzel M. E., "Perception of discourse boundaries and prominences in spontaneous Dutch speech," Lund University, Dept of Linguistics, *Working Papers*, no. 46, pp. 5–26, 1997.

[30] Watson D., Gibson E., "Intonational phrasing and constituency in language production and comprehension," *Studia Linguistica*, vol. 59, no. 2–3, pp. 279–300, 2005.