Prosodic boundaries in spontaneous Russian: perceptual annotation and automatic classification

Irina Nesterenko

Department of Phonetics, Saint-Petersburg State University
& Laboratoire Parole et Langage
Université de Provence
irina.nesterenko@lpl.univ-aix.fr

Abstract

Perceptual experiments with French and Russian speaking subjects were used to locate intonation phrase boundaries under different experimental conditions. Once inter-listener agreement had been evaluated, we built an automatic predictor based on human boundary/no-boundary judgments and then evaluated how well the predictor behaves. This predictor operates on acoustic features and we looked for an optimal combination of features to mimic perceptual experiment results.

1. Introduction

The issue of automatic prosodic annotation for speech corpora has become of great importance in recent years with the tendency to work with attested data and use large speech corpora to validate different linguistic hypotheses. We can speculate on different applications for such annotated data both in theoretical and applied studies, whatever their orientation: psycholinguistics, speech pathology or formal langage studies. Besides, in the speech synthesis domain, it was convincingly demonstrated that corpus driven unit selection techniques perform better when the speech database includes prosodic structure annotation [13, 14].

Our study deals with the issue of prosodic phrasing and its acoustic modelling in spontaneous Russian speech. We search for a model with perceptually justified levels of prosodic phrasing as well as for an acoustic model of prosodic boundaries with different strengths. It is generally agreed that prosodic phrasing is important in speech communication since it allows the listener to reconstruct the utterance's internal organization and informational structure, intended by the speaker. The perceived prosodic structure results from the complex acoustic-linguistic and cognitive processing. In speech, grouping seems to be the product of a trade-off between different interfaces, especially between prosody and syntax, on the one hand, and between prosody and discourse structure, on the other hand.

Once the perceptual listener-oriented approach is adopted (and we insist on it following the proposal by C. Wightman [15]), the matter of annotation accuracy arises. To evaluate this parameter, we adopt Kappa statistics for interrators' agreement which quantify the degree to which the common underlying annotation scheme is taken over by all judges.

Moreover, the Kappa coefficient is a useful measure when looking for the extent to which different acoustic features contribute to predict perceived boundaries. To test the impact of selected acoustic features, we run a series of experiments using a discriminant analysis. In fact, the output of such experiments is a contigency table resuming information on correct predictions and misclassification in comparison to initial human annotation. It has frequently been underlined that the proportion of correct classifications is not a good estimation of the model efficiency, given a large proportion of "no boundary" cases. Kappa statistics have rarely been applied in such studies though they allow to access classifier's effectiveness, by comparing it with a classification without any acoustic information.

The rest of this paper is organised as follows: section 2 presents the general project; we detail the perceptual experiments are detailed in section 3. Section 4 presents the results of inter-judge agreement with special attention paid to different statistics to be used in such analyses. The issue of acoustic correlates associated with perceived prosodic boundaries is further detailed in section 5. Finally, we discuss the impact of the approach adopted and required future work.

2. Project description

Our project aims at obtaining a multilevel annotation for a corpus of spontaneous Russian speech. First, we seek to formalise the concept of intonation unit: in the tradition of Russian intonational studies, this concept is interpreted in its close association with syntactic and semantic levels, its formals properties being relegated to a secondary plan. At the same time, Russian language has not benefited from a large volume of studies in prosodic phonology.

So, our first aim being an automatic corpus annotation in terms of intonational units, we decided to adopt a listener-oriented approach. The issue is then to propose an adequate methodology to assess subjects' judgments. In our study we opted for a meta-linguistic task proposed to trained phoneticians. This methodological paradigm is motivated in our case by the preceding analysis of task formulation used by different researchers: in several annotation studies (cf. [3]) a strong prosodic boundary is explicitely defined as one accompanied by a pause. Yet, it has been demonstrated in many studies carried in recent years that a prosodic boundary is a complex acoustic phenomenon associated with many indices. Being aware of all the underlying uncertainties and controversies associated with meta-linguistic tasks, different precautions were adopted as to experimental design and the interpretation of the results. We report first on perceptual judgments of intonational phrases boundary placement by French and Russian speaking subjects, while listening to Russian speech material.

3. Perceptual experiments

3.1. Corpus

In our study we choose to work on a corpus of spontaneous Russian speech. This choice is motivated, on the one hand, by the interest recently demontrated in the modeling of conversational speech in speech applications, as well as by the issue of variability in prosodic phrasing and closer relation manifested in unprepared speech between prosodic phrasing and informational structure.

The corpus of spontaneous dialogue speech in Russian was collected for the INTAS project 915 at the department of Phonetics, Saint-Petersburg State University. For the current study, the stimuli were selected from the recordings of an informal spontaneous dialogue between two female speakers in their twenties.

3.2. Stimuli

For the perception experiment we selected 25 inter-pausal units (IPU) of variable length from our corpus. Spontaneous speech is characterized by being less structured syntactically than prepared read speech, and contains many types of disfluencies and hesitation phenomena. We presume that such phenomena need special prosodic-acoustic signaling, hence, several of the stimuli chosen for the perceptual experiment contained such phenomena.

3.3. Subjects

Metalinguistic assessment of prosodic phrasing precludes addressing naïve speakers. Our research question being the extent to which prosodic phrasing is cued by prosodic features (in contrast to lexico-grammatical information), we decided to conduct the experiment with two groups of subjects. 7 Russian-speaking subjects and their 7 non Russian-speaking (in our case, French speaking) subjects took part in the experiment: all subjects were PhD students in the field of phonetics/prosody or faculty members of the speech labs.

3.4. Experiment

Being aware of all the pitfalls underlying meta-linguistic tasks, and especially the influence of the listener's conception of the task and of the associated concepts, we decided to introduce a Condition factor in our study. Simultaneously, we sought to elucidate the impact of different linguistic cues available to listeners in the assignment of a coherent prosodic structure to the stimuli. The impact of semantic-syntactic information was evaluated via the use of the interlanguage paradigm. At the same time, we tried to factor out different types of prosodic information (rhythmic as well as melodic cues) by asking the subjects to fulfill the task under three experimental conditions:

a) **Condition 1**: the stimuli were presented with flattened fundamental frequency (the flattening being obtained via PSOLA resynthesis with the value set to the mean F0 value for the stimulus); afterwards, a low pass filter was applied (threshold fixed at 500 Hz).

b) **Condition 2:** a delexicalised version of the stimuli was obtained by application of a low-pass filter (same threshold).

c) **Condition 3:** a natural sounding version was presented during the last session.

So, the experiment thus composed of three sessions for each subject. The task proposed to the subjects consisted in marking the intonational phrase boundaries, without any explicit definition of the concept being proposed. Each subject fulfilled the task individually. The listeners were asked to make judgements on the basis of acoustic cues only. They could listen to the speech stimulus or its parts as many times as they wished. The subjects navigated through the speech file displayed via the Praat program with a blank tier reserved for them to enter their segmentation.

4. Results: inter-judges agreement

In the analysis of the data obtained in the perceptual experiments, our main hypotheses are related to the influence of the Condition factor (which can diverge for French and Russian speakers) and with the inter- and intra-speaker variability in the judgements about appropriate prosodic phrasing.

4.1. Kappa statistics

Traditionally, one resorts to Kappa statistics when the question of measuring inter-annotator agreement arises. Both pairwise agreement and Kappa coefficients provide an estimate of the consistency in annotators' performance, though only the latter proceeds by comparison of the observed agreement with the probability of the two rators agreeing by chance. Mathematically this corresponds to:

$$K = \frac{p(A)-p(E)}{1-p(E)},$$

where *K* is the Kappa value, *p(A)* designates the proportion of observed agreement, and *p(E)* is the proportion of agreement that would have occurred by chance.

It should be noticed that reliability studies are very developed in the domaine of ToBI-style intonation transcription evaluation, though the measure of pairwise inter-rator agreement is priveledged there. Furthermore, two methods exist to calculate kappa statistics for multiple rators: a classical one proposed by Cohen [5] and the method proposed later by Siegel & Castellan [10]. The difference between two methods lies in that the traditional method does not assume equal classification proportions for the different rators. On the other hand, Siegel & Castellan's method provides an adjustment for bias, where the different rators systematically differ in their categorization. Given the annotation perspective, we consider the assumptions of the second method correspond better to the proposed task.

4.2. Agreement evaluation

We adopted the pair-wise method, as stricter than a mere comparison of the proportion of agreeing judges. However, the choice of the experimental paradigm and the quest of uniformity of data analysis did not allow us to use the classic "transcriber-pair-word" unit for this analysis (word division of the material was not accessible to the listeners under all conditions). So, we decided to include in our array of analysed units every syllable boundary, in order to dispose of a condition-independent set of analysed positions. In the case of delexicalised speech, our previous research confirmed the capacity of listeners to evaluate the length of a filtered stimulus in terms of number of syllables.

The following statistics were calculated:

Pairwise judge agreement was calculated for every syllable boundary position in three experimental conditions for French and Russian speaking subjects: the results are presented on Figure 1.

The data show that the mean pairwise agreement rate in our study is at the level of 90%. More marked fluctuations are induced by the Condition factor in the performance of Russian subjects as compared to French subjects, with a maximum reached for condition 3 (normal presentation of stimuli). This observation is validated statistically: linear regression models confirm a significant Condition effect for Russian speaking subjects ($F(2, 40) = 184.2$, $p < 0.0001$), but not for French subjects ($F(2,40) = 2.54$, $p = 0.0919$). However, as already mentioned, good pairwise agreement is largely due to "no boundary" cases, quantitatively dominent in the analysed material.

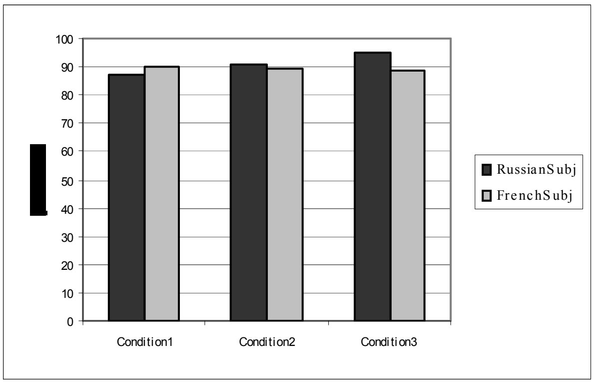

Figure 1: Pairwise agreement for the presence of prosodic boundary in three experimental conditions

For the calculation of **Cohen's Kappa statistics**, we used a build-in function in the R programm [6], which provides both Cohen's and Siegel & Castellan's coefficients as well as evaluating their statistical significance. All the values are significant at 0.01 level ($p << 0.00001$). The results from both methods are presented in Table 1.

Langue	Condition	Kappa Cohen	Kappa Siegel
Auditeurs russes	I	0,80	0,46
	II	0,84	0,54
	III	0,92	0,76
Auditeurs francophones	I	0,81	0,48
	II	0,83	0,50
	III	0,84	0,55

Table 1: Inter-auditor agreement evaluation with kappa statistics, two-method comparison

Discussion: Our first remark concerns the observed difference between the two measures: the values of the classic kappa coefficient are sometimes twice as large as those obtained with the second method. Though, we observed that the differences are slighter in the case of two rators: note, that some studies use these pairwise kappa values [3].

If we compare our results with those communicated by different researchers [11], the level of agreement achieved by Russian speaking subjects with normally presented stimuli is quite in line with those studies (we base our comparison on Siegel-Castellan statistics). Though the agreement level under other conditions as well as that observed in the performance of French speaking subjects are less high. We can attribute this particular status of normal presentation condition with Russian speaking subjects to the syntactically anchored tradition of the prosodic analysis of Russian: once the semantic-syntactic information is available to the listeners, the agreement is increased. The weaker agreement under delexicalised conditions suggests that the underlying acoustic model of prosodic phrasing is less established.

At the same time the obtained results corroborate those from psycholinguistic studies with event-related brain potential techniques [8], testing human sensivity to prosodic structure with nonsensical stimuli. These findings induced us to undertake further acoustic analyses in the search of acoustic parameters, which correlate with observed judgements on prosodic boundaries.

5. Acoustic correlates of perceived prosodic boundaries

Acoustic analyses were carried out on a subset of examined data: at this stage we decided to process "consensual" boundaries, i.e. those marked by four or more listeners in the perceptual study.

Different acoustic cues correlate with boundary perception; the presence of a silent pause and pre-boundary lengthening are the two most frequently cited in the literature [2]. They are also claimed to be necessary cues in speech applications. We did not share this conviction: for us a silent pause is neither a necessary nor a sufficient feature for boundary signaling in spontaneous speech. At the same time, our stimuli were selected in such a way to exclude the presence of any structurally motivated pause in them. We consequently decided to base our investigation on duration and tonal parameters alone.

All features were extracted from the corpus automatically by Praat scripts. A set of 34 indices were investigated, which could be organised in three groups:

a) durational measures were obtained at the level of syllable: we use z-score transforms, which allow to neutralise "intrinsic" effects [4];

b) total amplitude value (following [1]) that represents a quantified degree of perceived prominence;

c) tonal organization measures: this class is not homogeneous in itself.

For the tonal measures we worked from a stylized curve using the MoMel algorithme [7], i.e. interpolated over voiceless regions and with some microprosodic effects suppressed. First, we determined three spans in which measures are taken: we looked at two syllables before boundary region, two syllables after the boundary region, and what we call "potential untonation unit", i.e. a span ending at the analysed location and starting 0.853 seconds before (this value being the mean of non-terminal intonational unit duration in our perceptual experiment). For each of the spans we obtained raw measures of f0 minima, maxima, mean, standard deviation and f0 velocity. Two normalisation procedures were further applied to minimise the subject factor, logarithmic transformation followed by z-score normalisation. After that, different ratios were calculated to quantify boundary related phenomena: both boundary tones (via the ratios relating measures from a potential intonation unit and pre-boundary region) and resetting phenomena (via the comparison of pre- and post-boundary regions).

In the next step, acoustic measures were correlated with subjectif perceptual boundary phenomena, and this via a number of prediction experiments applying discriminant analysis technics. All the combinations of one, two and three acoustic parameters were tested for their predictive strength. Table 2 resumes the findings for the best parameter combinations for the 2nd and 3rd experimental conditions.

The acoustic parameters that give the best classifications vary under experimental condition and with the language factor. Though the most frequently returned parameter whose role seems quite important in the prediction of the perceived prosodic phrasing is the total amplitude: this finding points out at a strong association between prominence and boundary phenomena, reflected in different phonetic and phonological prosodic studies [8].

While presenting the discriminant analysis results we particularly insist on Kappa evaluation for the proportion of correctly predicted cases. Here the p(E) corresponds to the accuracy of classification in the absence of any acoustic information. The presented propotion allows us elucidate the impact of a combination of acoustic parameters chosen. We note that these corrected values are of the same magnitude as the kappa agreement data from our perceptual study.

6. General Discussion and Conclusions

From the beginning of our study we had two objectives: to collect information to build an acoustic model of the perceived prosodic phrasing and to discuss the role of kappa measures in speech studies.

Our perceptual study of French and Russian speakers' judgements about Intonational Phrase boundary placement reveals that listeners are able to carry out the prosodic phrasing annotation under the proposed experimental conditions. The methodology used differs from that adopted in ToBI style studies which aim to capture boundary strength after every word, thus dealing rather with the relative weighting of juncture phenomena. In our study, we opted for a metalinguistic paradigm: while aware of its disadvantages, we aimed to avoid some circularity and theory-dependence as to the number of levels in the prosodic hierarchy, an issue that requires further investigation for Russian. We are convinced as well that further speculation is needed on objective methods of assessing listeners' judgments on the acceptability of prosodic phrasing. At the same time, the results of the perception experiment have further implications for the issue of automatic prosodic annotation confirming the possibility of predicting the distribution of prosodic boundaries from input speech, i.e. under a non-lexical prosodic environment.

We also undertook an acoustic analysis seeking the best predictors of the perceived consensual prosodic boundaries. It seems that total amplitude is the best predictor, followed by f0 ratios reflecting pre-boundary phenomena. The revealed instrinsic tie between prominence and grouping phenomena should receive more attention in our future work.

As to evaluation measures, we show that Kappa statistics could be used in assessing both inter-listener agreement and predictive effectiveness of the classifying algorithm: based on these statistics, the conclusion seems plausible that the predictive algorithm has the performance comparable with that of human listeners.

Yet, the results presented here are limited to an acoustic domain. Our future work is aimed at extending the analyses with linguistic dimension: in particular, further exploration of the prosody-syntax interface is required for the model developed to be applicable in speech technologies. Jointly several theoretical issues need be elucidated as to the prosodic hierarchy in Russian.

7. References

1. Beckman, M. E. (1986). *Stress and Non-Stress Accent* (Netherlands Phonetic Archives No. 7).
2. Blaauw, E. (1995). *On the perceptual classification of read and spontaneous speech*. Doctoral dissertation, Utrecht University.
3. Buhman J., Caspers J., van Heuven V., Hoekstra H., Martens J.P., Swerts M. (2002). Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the spoken Dutch corpus. In *Proceedings of LREC 2002*, Las Palmas, May 2002, 779-785.
4. Campbell, N. (1992). *Multi-level timing in speech*, PhD thesis, University of Sussex.
5. Cohen, J. "A coefficient of agreement for nominal scales", *Educational and Psychological Measurement*, 20:37-46, 1960
6. Hirst, D.J., Di Cristo, A. & Espesser, R. (2000). Levels of representation and levels of analysis for intonation. in M. Horne (ed) *Prosody : Theory and Experiment,* Kluwer, Dordrecht.
7. Gussenhoven, C. & A.C.M. Rietveld (1992). Intonation contours, prosodic structure, and preboundary lenghthening. Journal of Phonetics, 20, 283-303.
8. Pannekamp, A. et al. (2005). Prosody-driven sentence processing: an Event-related Potential Study. *Journal of Cognitive Neuroscience*, vol. 17(3), 880-883.
9. Siegel, S. and Castellan, N.J.Jr. *Nonparametric statistics for the behavioral sciences*. Boston, MA: McGraw-Hill.
10. Syrdal, A.K. and McGory, J. (2000) "Inter-transcriber reliability of ToBI prosodic labelling", *Proceedings of ICSLP 2000*, vol. 3:235-238.
11. Taylor, P., and Black, A.W. (1999). Speech Synthesis by Phonological Structure Matching. In *Proceedings of Eurospeech-99*.
12. Van Santen, J. et al. (2005). Synthesis of prosody using multi-level unit sequences. *Speech Communication*, 46, 365-375.
13. Whiteman, C. (2002). ToBI or not ToBI? In *Proceedings of the First International Conference on Speech Prosody*, Aix en Provence April 2002.

Language	Condition	Parameters number	Correctly classified cases	Correctly classified cases, % corrected for by-chance agreement	F-measure for "no-boundary" category	F-measure for "presence of the boundary" category
Russian	III	1	73,40	46,81	0,747	0,718
		2	75,53	51,06	0,758	0,753
		3	77,66	55,32	0,784	0,769
	II	1	72.09	44.19	0,714	0,736
		2	72.09	44.19	0,714	0,736
		3	76.74	53.48	0,767	0,767
French	III	1	70.83	41.67	0,750	0,650
		2	71.88	43.75	0,727	0,710
		3	76.04	52.08	0,768	0,753
	II	1	71.11	42.22	0,724	0,698
		2	73.33	46.67	0,739	0,727
		3	77.78	55.56	0,787	0,767

Table 2. Discriminant analysis results