INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Eliciting meaningful units from speech

Daniil Kocharov[1], *Tatiana Kachkovskaia*[1], *Pavel Skrelin*[1]

[1]Saint Petersburg State University, Russia

kocharov@phonetics.pu.ru, kachkovskaia@phonetics.pu.ru, skrelin@phonetics.pu.ru

Abstract

Elicitation of information structure from speech is a crucial step in automatic speech understanding. In terms of both production and perception, we consider intonational phrase to be the basic meaningful unit of information structure in speech. The current paper presents a method of detecting these units in speech by processing both the recorded speech and its textual representation. Using syntactic information, we split text into small groups of words closely connected with each other. Assuming that intonational phrases are built from these small groups, we use acoustic information to reveal their actual boundaries. The procedure was initially developed for processing Russian speech, and we have achieved the best published results for this language with F_1 equal to 0.91. We assume that it may be adapted for other languages that have some amount of read speech resources, including under-resourced languages. For comparison we have evaluated it on English material (Boston University Radio Speech Corpus). Our results, F_1 of 0.76, are comparable with the top systems designed for English.

Index Terms: phonetics, prosody, syntax, automatic intonational phrases detection

1. Introduction

Mutual understanding between the speaker and the listener is a result of a complex process. A major role in this process belongs to phrasing—that is, organizing the speech flow into meaningful units, big enough to convey a piece of information, but at the same time small enough to be basic units used as building blocks forming the information structure of spoken text.

The task of automatic speech understanding relies heavily on detecting these meaningful units (intonational phrases). Each unit is made up of words connected closely with each other. In written text, these connections are realized by means of syntax and punctuation. In spoken text, punctuation is no longer relevant, and this is where prosody comes into play. Prosody not only replaces punctuation not present in spoken text, but also serves as the main means of organizing the speech flow: segmenting it into the basic meaningful units; ensuring each unit to be perceived as a whole; establishing connections between these units within larger ones.

The existing procedures developed for prosodic boundary detection differ in two major ways of predicting intonational phrases depending on the input: acoustic and textual. In case of text processing the commonly used features include: punctuation [1], estimation of phrase length [2], POS tags [2], [3], shallow constituent-based syntactic features [4], [1], deep syntactic tree features [5], [6] and word embeddings [7].

Different sets of acoustic features are used for prosodic boundary detection in speech. Common durational features include pause duration and various estimates of pre-boundary lengthening. These cues show high efficiency, and some systems are confined to durational features only [8] [9]. Features based on fundamental frequency include F_0 range, F_0 slope, mean F_0 and others [10] [11] [12]. Intensity, amplitude or energy are considered weaker cues [13], but are also used in some systems [11] [14].

The procedure presented in this paper combines the two sources—syntax and acoustics—in one system capable of predicting prosodic boundaries in speech. The paper includes descriptions of syntactic and acoustic components separately and in combination.

In practice, even in "regular", prepared speech syntactic and prosodic boundaries do not always coincide. Thus, a group of closely connected words can be split into two or more parts—due to pragmatic reasons, or when the phrase would be too long otherwise. This is why we have taken the following approach. We assumed that there are such word junctures where a prosodic boundary is highly *im*probable—e.g., between a preposition and its dependent noun. Based on this assumption, the syntactic component is designed to predict phrase boundaries with a recall close to 100 %; as a result, the text is split into short phrases—mostly 1–3 words long. At the next stage these phrase boundaries are used as input to the acoustic component: it chooses among only those word junctures where a syntactic boundary is possible.

The procedure was initially developed for Russian speech [15] and has shown excellent results. In order to prove its language-independence, we have applied it to English using Boston University Radio Speech Corpus.

2. Experimental speech corpus

2.1. Russian material

We are using Corpus of Professionally Read Speech (CORPRES) for evaluation of the presented method on the Russian speech [16]. It contains recordings of fictional and newspaper texts read by eight native speakers of Standard Russian, 4 males and 4 females. For our experiments we used the recordings of two fictional texts with about 35 000 running words recorded from eight speakers and one play with about 22 000 running words read by four speakers.

It contains approximately 26 hours of speech with manually produced and aligned prosodic, orthographic and phonetic transcription. Prosodic annotation includes boundaries of intonational phrases and prosodic words. Prosodic words carrying nuclear stress and additional prosodic prominence are marked.

2.2. English material

We are using Boston University Radio Speech Corpus [17] for comparative evaluation of our method. Since being published, BURNC is one of the main sources of experimental material to evaluate various methods of prosody processing. The corpus contains recordings from six speakers: 3 males and 3 females. The annotation contains manual orthographic and prosodic labeling and automatic phonetic labeling. The prosodic annota-

http://dx.doi.org/10.21437/Interspeech.2017-855

tion was performed in terms of ToBI system. In our experiments we use boundary labels "4" and higher as markers for boundaries of intonational phrases.

Since temporal markers play an important role in detection of prosodic units, the phonetic transcription and alignment need be perfect. The current phonetic annotation in the corpus did not fulfil our requirements. Thus, we have realigned the orthographic transcription with speech signal using Speech Science Web Services of Bavarian Archive for Speech Signals [18], [19], that gave us almost perfect alignment of speech with both orthographic and phonetic transcriptions. Prosodic annotation was further automatically aligned with the new orthographic speech segmentation.

3. Automatic detection of intonational units

The general procedure is language-independent, but the software implementation has some differences depending on the corpus design and its annotation format.

3.1. Acoustic features

One of the functions of prosody is to ensure that each unit is perceived as a whole. The major mechanism here is declination. It means that the beginning of a new utterance or intonational phrase is often marked by a reset of acoustic features: fundamental frequency, tempo, amplitude. In general, a new utterance or intonational phrase often begins with higher F_0, faster tempo and higher amplitude [20].

Our choice of acoustic features is largely based on the phenomenon of declination. For each prosodic word in the material we calculated the difference (reset) in F_0, tempo and amplitude between this word and the following word, assuming that a significant difference must be a marker of prosodic boundary. Besides these, we used a number of standard features (see below).

Pauses are known to be the most reliable cue for boundary detection. However, not all of the boundaries are marked in this way—it is not uncommon to observe a sequence of intonational phrases produced without pauses.

It is known that in many languages speakers tend to slow down towards the end of the utterance or intonational phrase; this phenomenon is often called pre-boundary lengthening [21], [22]. We estimate this slowdown by calculating tempo reset.

Estimation of pre-boundary lengthening is the only language-specific acoustic feature, since its scope and degree differs across languages. In Russian the main carriers of pre-boundary lengthening are stressed vowels [23], while for English it is the last rhyme of the word. Therefore, for our measurements we have chosen different target segments. For English, we analyzed the last vowel only, since the coda consonants are not always present.

For each prosodic word, estimation of pre-boundary lengthening is the difference between its target segment's normalized duration and that of the following word. For phone duration normalization, we used the following formula, which allows to compensate for the average duration of the segment and its standard deviation:

$$\tilde{d}(i) = \frac{d(i) - \mu_p}{\sigma_p} \quad (1)$$

where $\tilde{d}(i)$ is the normalized duration of segment i, $d(i)$ is its absolute duration, and μ_p and σ_p are the mean and standard deviation of the duration of the corresponding phone p. The mean and standard deviation are calculated over the whole corpus for each speaker separately.

The typical F_0 contour of a simple sentence is a sequence of rise-falls corresponding to prosodic words, where the phenomenon of declination is observed [20]. Mainly, F_0 declination implies the following: (1) F_0 maximum is located within the first word of the sentence; (2) F_0 range decreases towards the end of the sentence.

Another phenomenon concerns the F_0 contour within the nucleus. The location of nuclear stress, where a major F_0 change is often observed, is language-dependent. Our experience and corpus data show that it is often located on the last prosodic word within the intonational phrase.

Based on these tendencies, we are using two features:

- F_0 reset (in semitones)—for each prosodic word, it is the difference between its maximum F_0 value and that of the following word;
- F_0 range (in semitones) within the prosodic word;
- F_0 maximum (in semitones) within the prosodic word;
- cumulative F_0 movement (in semitones) over the whole prosodic word;
- the steepest angle of the F_0 curve within the prosodic word.

The declination of F_0 contour if often matched with the declination of intensity. We estimate this reset using energy values: for each prosodic word, it is the ratio between its maximum energy and that of the following word.

3.2. Syntactic features

Splitting text into syntactic phrases requires morphological tagging and syntactic parsing. So far we are applying different parsers for English and Russian. Syntactic parsing of Russian texts was performed using Solarix [24]—a free tool with built-in lexicons and pre-trained models. Syntactic parsing of English texts was performed using SyntaxNet, a TensorFlow implementation of Parsey McParseface [25]—an open-source syntactic parser based on a feed-forward multi-layer perceptron. In our experiments we have used a pre-trained model for English provided by SyntaxNet. Both parsers accept plain text as input, split it into individual sentences and tokens, perform probabilistic part-of-speech tagging, and build a dependency tree by choosing the most probable type of syntactic link for each pair of words.

Our procedure of splitting text into a sequence of syntactic phrases based on dependency trees was developed analytically. For English we have used the same set of rules as for Russian.

The principal rule is based on the assumption that prosodic boundaries may appear only at those word junctures where syntactic discontinuity is observed.

In order to improve syntactic phrasing prediction, a set of post-hoc rules was applied:

1. Prepositions and articles. In cases of syntactic discontinuity between the preposition/article and the following word, the preposition/article was automatically assigned to the following group: e.g. [a] [new report says] → [a new report says]. In this example the article "a" is connected with "report", and then only the noun "report" governs the adjective "new"; thus, in terms of syntactic trees, there is no direct connection between "a" and "new". Introduction of this rule enables to deal with such cases.

2. Conjunctions. Since a boundary is quite possible before a coordinating conjunction, it was separated from the previous word even if no discontinuity was observed: e.g. [white][red and blue stripes] → [white][red][and blue stripes].

3. Splitting long nested phrases. Long phrases without syntactic discontinuity were divided further using the following principle: a boundary was placed before a word if its parent stood next to it, but formed its own group with another word, e.g. long noun phrase [the development of algorithms for solving equations] → [the development of algorithms][for solving equations] or long verb phrase [the current leader reaches the retirement age] → [the current leader][reaches the retirement age]. This rule was added to account for the limitations on IP length: when a phrase is too long, speakers tend to split it into smaller parts, and these parts are not random—breaks are inserted at those junctures where the connection is weaker.

Boundaries between intonational phrases fall on syntactic boundaries in about 97 % of cases for both languages. The linguistic analysis of the rest 3 % shows that there are language-specific rules that could deal with these cases. For each language there is a set of approximately 10–20 rules that can add 2–2.5 %. They include rules concerning parentheses, phrasal verbs, compound numbers, double names, collocations, and so on.

Of course, some of the errors in IP boundary detection are caused by parsing errors. Unfortunately, we are unable to calculate the parsing accuracy, since the "golden standard" syntactic annotation for our material is not available yet.

Thus, we work on with the initial design providing a 0.97 recall, losing 3 % of boundaries. This shift of performance was accounted for in the evaluation of overall performance presented in Tables 1 and 2.

The 3 % loss is more than compensated by a boost in precision; this boost is language-specific, or it may be corpus-specific depending on the annotation scheme. For Russian the precision increased from 0.37 to 0.61, i. e. two thirds of non-final words have now been correctly detected as non-final. For English the precision increased from 0.19 to 0.33, i. e. more than half of non-final words have now been correctly detected as non-final.

3.3. Combining syntactic and acoustic features

Working with texts, we can only speak of predicting the information structure of corresponding speech. The same utterance may be realized by different speakers in different ways with no significant change in meaning [26]. Moreover, different speakers may *understand* the text differently and, as a consequence, split it into prosodic units in different ways. This may be illustrated by an example from CORPRES corpus, where the phrase "along a blind long stone fence" produced by eight speakers was never pronounced as one IP: five speakers split it after the first adjective ([along a blind][long stone fence]), and three speakers—after the second ([along a blind long] [stone fence]). These realizations do not differ functionally or semantically, and listeners perceive both as neutral.

This is why we propose a two-stage procedure of combining syntax and acoustics. The first stage relies on syntactic data and consists in predicting all possible prosodic boundaries based on text. In other words, this step eliminates those junctures where a boundary is virtually impossible. Now, only the probable junctures are passed on to the next stage.

At the second stage acoustics come into play: using a statistical classifier, we perform automatic classification of word junctures predicted at the first stage based on our set of acoustic features. The classification is performed by means of Random Forests classifier [27]. The set of prosodic words is considered a homogeneous set of individual units, but not an organized sequence of units.

Table 1: *Performance of various prediction setups for Russian*

Feature set	Pr	Rec	F_1	Acc (%)
Acoustic features	0.944	0.785	0.857	90.2
Syntactic info and acoustic features	0.993	0.852	**0.917**	92.0

Table 2: *Performance of various prediction setups for English (BURNC): the method described in this paper, and state-of-the-art unsupervised [28] and supervised [29] systems*

Feature set	Pr	Rec	F_1	Acc (%)
Acoustic features	0.862	0.618	0.72	91.1
Syntactic info and acoustic features	0.862	0.682	**0.762**	86.5
Baselines				
Acoustic features, unsupervied [28]	0.80	0.65	0.73	85.7
Acoustic features, supervied [29]			0.736	90.7
Syntactic info and acoustic features, supervied [29]			0.761	91.5

4. Results

To assess our procedures we use precision (Pr), recall (R) and F_1 measure. Precision is the number of correctly predicted boundaries relative to the total number of predicted boundaries. Recall is the number of correctly predicted boundaries relative to the total number of boundaries in the corpus. F_1 measure is calculated using the following formula:

$$F_1 = 2 \cdot \frac{Pr \cdot R}{Pr + R} \tag{2}$$

In our case it is unreasonable to evaluate the performance by means of accuracy, which takes into account type I errors for both classes, as the number of final words is several times larger than that of non-final words. Thus, if we label the whole text as one large phrase with no informational structure and no meaningful units at all, the overall efficiency for the English material would be 81 %. Statistically, this result seems quite a success, but linguistically it is clearly absurd. Still we provide accuracy here for comparison with other published results.

The experimental material from both corpora was divided into training set (70 %) and test set (30 %). The speaker-dependent data were equally distributed among these sets.

Table 1 presents the efficiency of the method for Russian. After adding the syntactic component, there is a clear boost in performance. In the end we obtain the F_1-measure of 0.917, which is the highest for Russian so far.

For comparison, and in order to test our procedure on other

languages as well, we have adapted it for Boston University Radio Speech Corpus.

Table 2 provides our performance data compared with state-of-the-art systems. All of it was evaluated on the same data. Using only acoustic features, we have almost reached the efficiency of both the competing methods. When it comes to combining syntax and acoustics, our results are even slightly better, but the difference is insignificant.

5. Conclusions

The presented two-stage procedure for automatic prosodic boundary detection has shown high efficiency. There is a significant increase in performance compared with the acoustically-based setup. In terms of F_1 measure, it is around 6 % for both languages:

- English: $0.72 \rightarrow 0.762$,
- Russian: $0.85 \rightarrow 0.917$.

We assume that the remarkable difference in efficiency between Russian and English data is largely due to the differences in speech material. The Russian corpus (CORPRES) contains recordings of fictional texts, where syntax and punctuation convey most of the information so that the reader could understand it. During the recording process, the task of the speaker is to "deliver" this information to the listener in a very clear way. As a result, most boundaries are strongly marked, which is easy to recognize for a machine. The English corpus (BURNC) contains spontaneous recordings, where part of the information is conveyed by prosody only. Planning difficulties lead to unpredictable syntax-prosody relations: splitting closely connected words, joining loosely connected ones, truncating phrases for self-corrections etc. This reduces the performance of the automatic procedure.

There are language-independent features for signaling prosodic units in speech [20]. This is why the set of acoustic features applied for the task of detecting these units by different researchers is more or less the same. It includes various measures of within-word and across-word phenomena based on calculating maximum and minimum values, dynamics and resets of F0 and energy, and various measures of segmental duration. The actual sets of features vary from research to research.

We use a language-independent principle for syntactic prediction of prosodic boundaries formulated as follows: The perceived prosodic discontinuity which reflects the informational structure of utterance is intentionally realized by a speaker in places of syntactic discontinuity.

This principle is easy to implement as long as a syntactic parser for a given language is available. One can argue that the high-level parser requires a lot of syntactically labeled data, and that is true. However, the number of syntactic resources is increasing. Thus, for example, the Universal Dependencies Treebank Collection [30] now contains 50 languages including under-resourced, such as Estonian, Catalan, Galician and many more. There are language-independent parsing engines that use such resources as MaltParser [31] and SyntaxNet.

No linguistic expertise is required to replicate our syntactic processing method for recognizing meaningful units in speech in other languages.

6. Acknowledgments

The acoustic modeling and Russian material processing was developed as a part of research grant "Automatic segmentation of speech into prosodic units" supported by the Russian Science Foundation (# 14-18-01352). We thank Martti Vainio for the possibility to work with Boston University Radio Speech Corpus at the Institute of Behavioural Sciences, University of Helsinki.

7. References

[1] O. Khomitsevich and P. Chistikov, "Using statistical methods for prosodic boundary detection and break duration prediction in a russian tts system," in *Proceedings of Dialogue 2013*, 2013, vol. 2, pp. 11–19.

[2] P. Taylor and A. W. Black, "Assigning phrase breaks from part-of-speech sequences," *Computer Speech & Language*, vol. 12, no. 2, pp. 99–117, Apr. 1998.

[3] I. Read and S. Cox, "Using part-of-speech tags for predicting phrase breaks," in *Proceedings of Interspeech 2004*, Jeju Island, Korea, Oct. 2004, pp. 741–744.

[4] B. Lobanov, "An algorithm of the text segmentation on syntactic syntagrams for TTS synthesis," *Proceedings of Dialogue 2008*, 2008.

[5] J. Hirschberg and O. Rambow, "Learning prosodic features using a tree representation," *Proceedings of Eurospeech 2001*, pp. 1175–1178, 2001.

[6] S. Hoffmann, "A data-driven model for the generation of prosody from syntactic sentence structures," Ph.D. dissertation, ETH-Zürich, Zürich, 2014.

[7] A. Vadapalli and S. V. Gangashetty, "An investigation of recurrent neural network architectures using word embeddings for phrase break prediction," in *Interspeech 2016*, 2016, pp. 2308–2312.

[8] T.-j. Yoon, J. Cole, and M. Hasegawa-Johnson, "On the edge: Acoustic cues to layered prosodic domains," in *Proceedings of ICPhS'2007*, Saarbrücken, Germany, 2007, pp. 1264–1267.

[9] C. W. Wightman and M. Ostendorf, "Automatic recognition of prosodic phrases," in *Proceedings of ICASSP-91*, 1991, pp. 321–324 vol.1.

[10] N. Segal and K. Bartkova, "Prosodic structure representation for boundary detection in spontaneous French," in *Proceedings of ICPhS'2007*, 2007, pp. 1197–1200.

[11] J. H. Jeon and Y. Liu, "Semi-supervised learning for automatic prosodic event detection using co-training algorithm," ser. ACL '09, vol. 2. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 540–548.

[12] U. Reichel and K. Mády, "Comparing parameterizations of pitch register and its discontinuities at prosodic boundaries for Hungarian," in *Proc. Interspeech 2014*, Singapore, 2014, pp. 111–115.

[13] L. Streeter, "Acoustic determinants of phrase boundary perception," *The Journal of the Acoustical Society of America*, vol. 64, no. 6, pp. 1582–1592, 1978.

[14] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1526–1540, Sep. 2006.

[15] D. Kocharov, T. Kachkovskaia, A. Mirzagitova, and P. Skrelin, "Combining syntactic and acoustic features for prosodic boundary detection in russian," in *Proceedings of the Statistical Language and Speech Processing: 4th International Conference*. Springer International Publishing, 2016, pp. 68–79.

[16] P. Skrelin, N. Volskaya, D. Kocharov, K. Evgrafova, O. Glotova, and V. Evdokimova, "Corpres - corpus of Russian professionally read speech," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, P. Sojka, A. Hork, I. Kopeek, and K. Pala, Eds. Springer Berlin Heidelberg, Sep. 2010, no. 6231, pp. 392–399.

[17] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, *The Boston University Radio News Corpus*, 1995, tech. report.

[18] F. Schiel, "Automatic phonetic transcription of non-prompted speech," in *Proc. of the ICPhS*, San Francisco, August 1999, pp. 607–610.

[19] T. Kisler, U. Reichel, F. Schiel, C. Draxler, B. Jackl, and N. Prner, "Bas speech science web services – an update of current developments," in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), 2016.

[20] J. Vaissière, "Language-independent prosodic features," in *Prosody: Models and Measurements*, ser. Springer Series in Language and Communication, A. Cutler and D. R. Ladd, Eds. Springer Berlin Heidelberg, Jan. 1983, no. 14, pp. 53–66.

[21] W. E. Cooper and M. Danly, "Segmental and temporal aspects of utterance-final lengthening," *Phonetica*, vol. 38, no. 1-3, pp. 106–115, 1981.

[22] D. Byrd, J. Krivokapic, and S. Lee, "How far, how long: On the temporal scope of prosodic boundary effects," *The Journal of the Acoustical Society of America*, vol. 120, no. 3, pp. 1589–1599, 2006.

[23] T. Kachkovskaia, "The influence of boundary depth on phrase-final lengthening in Russian," in *Statistical Language and Speech Processing*, ser. Lecture Notes in Computer Science, A.-H. Dediu, C. Martn-Vide, and K. Vicsi, Eds. Springer International Publishing, Nov. 2015, no. 9449, pp. 135–142.

[24] "Solarix," 2016, http://www.solarix.ru.

[25] D. Andor, C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins, "Globally normalized transition-based neural networks," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, August 2016, pp. 2442–2452.

[26] S. R. Speer, P. Warren, and A. J. Schafer, "Situationally independent prosodic phrasing," *Laboratory Phonology*, vol. 2, no. 1, pp. 35–98, 2011.

[27] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[28] A. Suni, J. imko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech & Language*, in Press.

[29] A. Rosenberg, "Automatic detection and classification of prosodic events," Ph.D. dissertation, Columbia University, 2009.

[30] J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman, "Universal dependencies v1: A multilingual treebank collection," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

[31] J. Nivre, J. Hall, and J. Nilsson, "Maltparser: A data-driven parser-generator for dependency parsing," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, 2006.