PAN ARCHIVES OF ACOUSTICS
Vol. 41, No. 2, pp. 233–243 (2016)

DOI: 10.1515/aoa-2016-0023

Detection of Sentence Boundaries in Polish Based on Acoustic Cues

Magdalena IGRAS[1], Bartosz ZIÓŁKO[1,2]

[1] *Department of Computer Science, Electronics and Telecommunications*
AGH University of Science and Technology
Al. Adama Mickiewicza 30, 30-059 Kraków, Poland; e-mail: {migras, bziolko}@agh.edu.pl

[2] *Techmo sp. z o.o.*
Poland techmo.pl

(*received November 27, 2015; accepted January 15, 2016*)

In this article the authors investigated and presented the experiments on the sentence boundaries annotation from Polish speech using acoustic cues as a source of information. The main result of the investigation is an algorithm for detection of the syntactic boundaries appearing in the places of punctuation marks. In the first stage, the algorithm detects pauses and divides a speech signal into segments. In the second stage, it verifies the configuration of acoustic features and puts hypotheses of the positions of punctuation marks. Classification is performed with parameters describing phone duration and energy, speaking rate, fundamental frequency contours and frequency bands. The best results were achieved for Naive Bayes classifier. The efficiency of the algorithm is 52% precision and 98% recall. Another significant outcome of the research is statistical models of acoustic cues correlated with punctuation in spoken Polish.

Keywords: punctuation; sentence boundary; spoken language; prosody; Polish.

1. Introduction

The main motivation of the research is to improve automatic speech recognition systems with automatic sentence boundary detection by exploiting information from acoustic or prosodic cues.

In existing solutions all over the world, it is necessary to verbalize punctuation commands (for example: 'full stop', 'comma'). The automatic speech recognition (ASR) systems for Polish (Ziółko *et al.*, 2011) create direct speech transcripts, without upper-cases or punctuation. It decreases user comfort and makes the transcripts less easily readable. At the same time, the disfluencies (e.g. filled pauses, restarts or repairs) present in natural spontaneous speech interfere with sentence borders (Barczewska, Igras, 2013; Igras, Ziółko, 2013a).

The indicators of punctuation are present in speech and intuitively perceived by listener, but the information is lost during standard speech to text processing. By exploiting them, not only direct users will benefit, but it will also bring additional information for improving current systems. The majority of the state-of-art works for Polish (see Subsec. 2.2) investigate phenomena connected to intonational (prosodic) phrases but there is a lack of a complex study that would define indicators of punctuation marks. We focused our work on sentence or clauses boundaries (denoted with punctuation marks), not the prosodic boundaries. According to our knowledge, no other algorithm of automatic punctuation prediction for spoken Polish was introduced, nor for an acoustic prosody model or a language model.

The goal of this research was to find an algorithm that locates sentence or clauses boundaries on the basis of the acoustic information. We assume that acoustic or prosodic cues are natural discourse demarcation indicators, complementary to speech content. After finding proper acoustic correlates of punctuation marks we should be able to recognize majority of syntactic boundaries without knowledge on lexical content of the speech.

The paper is organized as follows: Sec. 2 introduces the background of the punctuation correlates in speech, taking into account previous works for Polish and different languages. Next, the database is described in Sec. 3. In Sec. 4 we summarize several groups of acoustic features and their connection with location in a sentence. In Secs. 5–7 statistical and classification

tools as well as our experiments are described. The results are discussed in Sec. 8 and the paper is concluded in Sec. 9.

2. Background

Prosodic cues convey structural, semantic and functional information. Most of them are resistant to communication channel characteristics changes. It was showed that over 65% of syntactic boundaries were coded in prosodic information (Fach, 1999). Prosodic parameters similar to those correlated with sentence breaks were found to indicate also larger breaks such as topics (Shriberg *et al.*, 2000). It was suggested that the natural unit of speech is the phrase rather than the sentence (Gotoh, Renals, 2000). Several experiments (Beeferman, Lafferty, 1998; Stevenson, Gaizauskas, 2000) showed that humans do not always agree on the insertion of sentence boundary or comma into transcriptions.

The outcome of the undertaken research will contribute to the following fields of speech and language technology:

- Speech modeling on segmental and suprasegmental level:

  The acoustic features of the same phoneme can differ for many reasons (neighborhood of other phonemes, speaker individual features emotional state, speech pathology, the accentuation and finally the phoneme position in a sentence). Therefore, understanding the location-dependent changes in phoneme characteristics will be useful for language modeling, e.g. for ASR.

- Computational linguistics:

  Both lack of punctuation and occurrence of disfluencies in spontaneous speech transcripts are factors that disturb their processing by natural language processing (NLP) systems. Automatic segmentation of the speech into phrases will adapt them to be processed by language models in large scale NLP.

- Descriptive linguistics:

  Modeling acoustic correlates of punctuation can bring additional arguments to discussion on language-dependent nature of punctuation and discourse analysis (see Subsec. 2.1).

- Speaker biometry:

  Connotations between pauses and punctuation as well as frequency and types of pauses vary between individuals. Thereby, the temporal features can be used for speaker biometry, speakers' diarization or evaluation of speaker oratorical skills (Barczewska, Igras, 2013; Igras, Ziółko, 2013a).

- Speech synthesis:

  More natural sounding can be achieved thanks to proper predicting of the positions and durations of prosodic breaks on the basis of syntactic structure (Navas *et al.*, 2008; Chistikov, Khomitsevich, 2013).

2.1. Related work for other languages

Simultaneously with advancement of speech technology systems, the task of sentence boundaries detection gained attention initially in the area of natural language processing (e.g. Beeferman, Lafferty, 1998). Then (about 2000) the efforts included speech domain, especially since prosody was proven to be less sensitive to speech recognition errors (Shriberg *et al.*, 2000). Still, the indicators of sentences boundaries seem to be language-dependent. While acoustic correlates of phrases were investigated in details for other languages and several classification methods were developed, for Polish such a complex tool has not been proposed yet.

The majority of approaches to punctuation annotation for other languages combined two models: lexical (usually n-grams) and prosodic model (containing features sets describing pitch, energy, pauses and phonemes length) (Shriberg *et al.*, 2000; Gotoh, Renals, 2000; Baron *et al.*, 2002). In many cases the latter were argued to be more robust than the lexical ones (Shriberg *et al.*, 2000; Baron *et al.*, 2002; Kim, Woodland, 2003). There were also studies that did not include language model, e.g. instead of speech recognition, only vowel/consonant/pause distinction was performed (Wang *et el.*, 2003). The prosodic parameters were usually extracted automatically by forced alignment with transcriptions or recognized words (Shriberg *et al.*, 2000), with no human labeling (direct prosody modeling – Shriberg, Stolcke, 2004).

The region of interest was usually a word (or 200 ms window) before and after a break. The most frequently exploited classifiers were: CART-style binary trees (Shriberg *et al.*, 2000; Kim, Woodland, 2003), neural networks (Christensen *et al.*, 2001). Good results were also achieved with maximum entropy approach (Huang, Zweig, 2002) and finite state machine models – FSM (Gotoh, Renals, 2000). For combining the models, hidden Markov models (HMM) were usually applied. It was also observed that the cues may be corpus-dependent: pause and pitch features were argued to be highly informative for segmenting news speech, whereas pause, duration and word-based cues dominated for natural conversation (Shriberg *et al.*, 2000). It was also concluded that there is a large difference in the usability of the different features for the current task and a large variation in how much discriminant information the prosodic

Table 1. Comparison of research using prosodic or acoustic parameters for sentence boundary detection.

Paper	Material	Parameters	Model	Lexical cues	Efficiency
Gotoh, Renals, 2000	English Broadcast News	pauses duration	finite state model	+	f-measure: 70%
Shriberg *et al.*, 2000	English Broadcast News and Switchboard	pause durations, phone durations, pitch, voice quality	CART, HMM	+	SER: 0–25%
Kim, Woodland, 2003	English Broadcast News	pause duration, F0, energy	CART	+	f-measure: 76%
Christensen *et al.*, 2001	English Broadcast News	pause duration, phone duration, F0	finite state model, MLP	+	f-measure: 42%
Baron *et al.*, 2002	English, multi-party meetings	energy, duration, F0	CART, HMM	+	Acc: 81–92%
Wang *et el.*, 2003	English, Broadcast news	pauses, F0, speaking rate	AdaBoost	–	Acc: 82–87%
Kolar *et al.*, 2004	Czech, Broadcast news	F0, speaking rate, loudness	CART, MLP	+	Acc: 95.2%
Wang, Narayanan, 2004	Switchboard	pitch breaks and durations	Linear fold algorithm	–	Acc: 75%
Vicsi, Szaszak, 2006	Finnish, Hungarian, read speech	F0, energy	HMM	–	Corr: 69–77%
Batista *et al.*, 2008	Portuguese, Broadcast news	acoustic segment changes, pauses	ME	–	SER: 74%

information carry (Christensen *et al.*, 2001). Another approach, using curve fitting to find pitch breaks correlated with sentence boundaries, was proposed in (Wang, Narayanan, 2004). A comparison of the mentioned works is collected in Table 1.

2.2. Review of research on Polish punctuation acoustic correlates

Polish, as a Slavic language, is characterized by relatively free words order and a complex inflectional morphology. Polish punctuation system is described as mainly syntactic (enhancing logic structure of utterance), still as an additional factor, an influence of rhythmics and intonation of speech is also included (contemporary dictionaries, e.g. *PWN Dictionary, 2013*). Likewise, authors of recent elaborations on modern Polish emphasize logical-syntactic role of punctuation, and as the secondary aspect the prosody (intonation, pauses, breathing) is taken into account (Karpowicz, 2012). Therefore, the role of the proper punctuation usage covers: exact expression of an author intention, suggesting intonation for a reader, indicating places where a pause should occur.

Historically, Polish writing system was inherited from Medieval Latin. Similarly to Latin, in the earliest Polish writings points were used to mark breaks: full stop for a long pause, colon for a medium and comma for the shortest one, in order to facilitate the reader using pauses (Przyłubski, 1953). Punctuation rules were codified in XVIII century focusing on semantic structure, and also speech rhythm or emotional tone. Evolving, language seemed to lose the initially strict connection with prosody towards mirroring semantic and syntactic structure. Hence, punctuation marks are not direct analogues of speech phonetics but rather a tool to organize content in order to make the text functional and usable (Łuczyński, 1999).

In contemporary Polish punctuation, there are more rigid rules concerning the use of commas (dependent clauses always being set off with commas, and commas being generally proscribed before certain coordinating conjunctions) and they appear more frequently than in English. In the realm of language prosody, it has to be marked that intonation of a language consists of some universal and some language-specific features. Polish accentual system has a regular word stress usually on the penultimate syllable of the word. There is no fixed sentence accent. The main function of the accent is to mark the word center and it does not serve as a word demarcation indicator. Vowels duration makes no semantic distinction and it is longer on accents (Ostaszewska, Tambor, 2000). The declarative sentences are rather flat, monotonous and a regular fall is noticeable on the last prominent word (Igras, Ziółko, 2014b). In both languages, the distribution of intonation patterns is affected by lexical and syntactic structure of the

text, but for Polish, high pitch variability is usually assigned to emotional speech.

Prosody has been a subject of study of many phoneticians, beginning from the pioneer works by Jassem (1962), Dłuska (1976), Steffen-Batóg (1996). Some recent works on Polish prosody were provided by Klessa *et al.* (2002; 2008; 2011), Karpiński *et al.* (2002), Wagner *et al.* (2010; 2011), Malisz and Wagner (2012), Wypych (2011) and Demenko (1999). In the latter, not only statistical analysis was introduced, but also automatic classification of intonational phrase structure as well as accentual structures were performed. The models of suprasegmentals were applied to speech synthesis. The detailed analysis of correlations between place in a phrase and phonemes duration influenced by accent and intonational context was also described in (Demenko, 1999). The research showed that, in Polish, the last and the one before the last vowel in a sentence is lengthened regardless of accent presence, but the lengthening effect is stronger for accented syllables (Jassem, 1973). Phrase perception depending on phoneme duration was studied on logatomes (Frąckowiak-Richter, 1973). The dependence of segmental duration on a selection of interacting factors is discussed by Klessa (2011). In another paper she finds out that the duration of particular syllable parts depends significantly on the presence of word stress and the syllables position with respect to pauses (Klessa, Śledziński, 2008).

A study of intonation phrases' boundaries was conducted by Wagner and Demenko (2007), Wagner (2010; 2011), Wagner *et al.* (2015) with focus on text-to-speech (TTS) synthesis systems. Prosodic boundaries were automatically classified with respect to strength (major or minor) and type (falling and rising) using artificial neural networks, discriminant function analysis and decision trees, with average accuracy between 79 and 82% (Wagner, 2010).

3. Materials

3.1. Training set

For an analytical study two corpora were used. First, pauses and statistics of phrases duration in spontaneous speech were investigated using a set of monologues in Polish. The prepared corpus consisted of different types of monologues (public, presentations, orations, real time translation, speeches and reports from the European Parliament from (Lööf *et al.*, 2009) in formal or half-formal situations, by experienced or professional speakers (politicians, professors, professional translators) and inexperienced speakers (students). The types of speech were chosen to ensure the diversity of utterance styles. Total duration of recordings is 60 min for 24 speakers. Positions of punctuation marks in recordings' transcriptions were annotated in MLF standard.

For modeling the suprasegmental features of phrases, Polish speech database CORPORA (Grocholewski, 1997) was used. The utilized part of the database contains 114 short sentences, each spoken by 45 speakers (males and females at different age). A part of the database was manually segmented into phonemes and the rest was segmented automatically and manually checked afterwards. Finally, 5130 observations of sentences' ends were extracted from the corpus.

3.2. Testing set

As a testing set, 15 minutes of speech (for 3 speakers, different from that in the training set) were annotated into phonemes and the punctuation marks were included in the annotation. Recordings from AGH Audio-Visual Speech Database (Igras *et al.*, 2012) and one recording of prepared monologue were used. Finally, the testing set contained 159 punctuation marks (98 full stops and 61 commas). All the utterances contained informative, continuous speech.

An example part of an MLF file enriched with punctuation annotation is presented below.

```
#!MLF!#
"D:/ANOTACJA/do anotacji/MIG/MIG.wav"
18780000 20000000 s
20000000 20640000 t
20670000 21480000 u
21480000 21770000 d
21770000 22260000 j
22260000 22910000 a
(...)
30960000 31710000 v
31710000 32100000 j
32130000 32530000 e
33080000 34030000 6
34030000 34950000 y , (comma)
35490000 36150000 k
36150000 36890000 t
36900000 37080000 u
37080000 37340000 r
37430000 37850000 o
(...)
54380000 54920000 z
54920000 55670000 a
55670000 56150000 v
56150000 56860000 o
56860000 57580000 d
57580000 58450000 o
58450000 59020000 v
59020000 59590000 e
59590000 60450000 j . (full stop)
68460000 68640000 t
68640000 69150000 o.
```

3.3. Data preparation

In the beginning, the amplitude of each utterance was normalized. Then, an algorithm of silence detection was applied in order to divide a recording into phrases, taking into account statistics of phrases in spontaneous speech (see Subsec. 4.1, 4.2). Finally, the extracted phrases are parts of continuous speech (of minimal length of 1 s and maximal length of 15 s) divided by a silent pause of minimal length of 200 ms. An example of the phrase extraction is presented in Fig. 1. The primary observation windows are frames 500 ms long. The choice of window length corresponds to the average duration of two syllables. The way of dividing the phrases into 0.5 s frames with overlap of 100 ms is shown in Fig. 2.

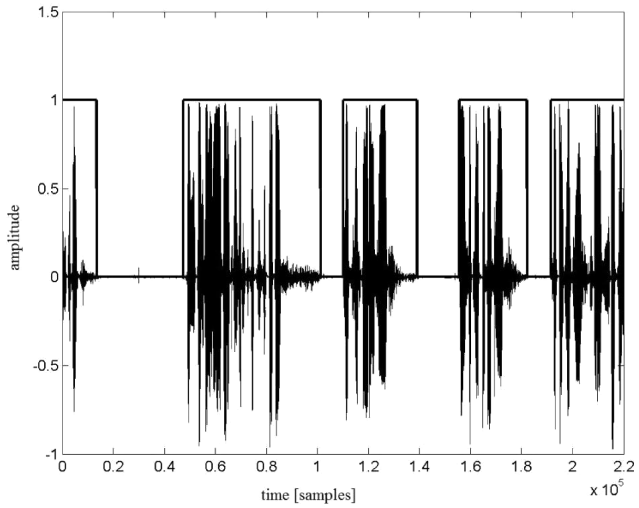


Fig. 1. Segmentation of speech signal into phrases separated by pauses.

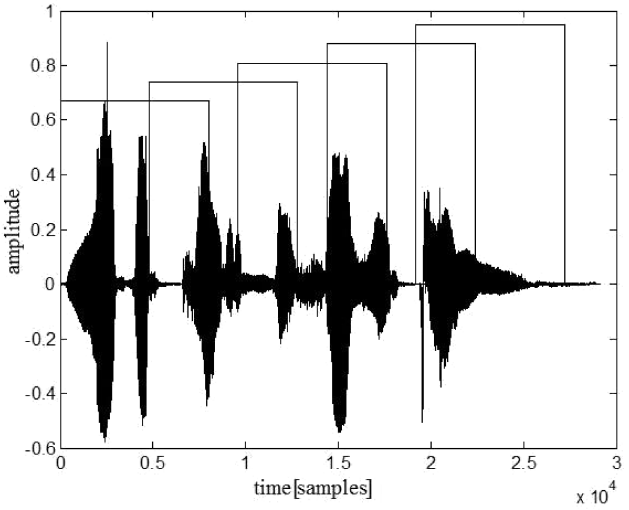


Fig. 2. Division of speech signal into frames of 500 ms.

On the basis of transcriptions, each frame was labeled with tags: <start> for the first frame in the sentence, <fstop> for the last frame before a full stop and <non> for the rest of the frames.

4. Features and models

4.1. Phrases length

Duration of a phrase of speech can be dependent on breathing rhythms, sentence syntax and meaning or individual speaker habits. Statistics of phrases length were collected from one hour of monologues. As we estimated, speech rate in spontaneous monologues is about 115 words per minute (with standard deviation between speakers of about 20 words/min).

Mean length of a sentence (containing on average 19 words) was about 10 seconds, while mean length of a speech unit separated with a punctuation mark (average 7 words) – 3.8 s. The results were similar for both orations or presentations and real time translations (Igras, Ziółko, 2013a; 2014b).

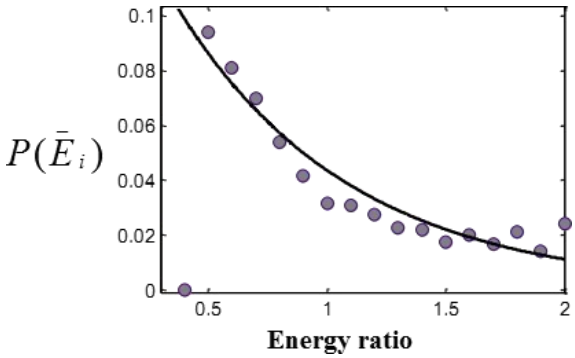
4.2. Pauses

Pauses are considered as the strongest indicators of punctuation. Using the corpus of monologues, we investigated also connotations of different types of pauses (silent pauses, breath pauses and filled pauses) with full stops and commas (Igras, Ziółko, 2013a). 39% of all full stops are connotated with occurrences of a breath pause, 27% with a silent pause, 20% with a filled pause. 28% of all commas are pointed by a silent pause, 20% by a breath pause and 6% by a filled pause. Lack of any kind of pause (words bonding in pronunciation) was registered in 20% of full stops occurrences and 46% of commas for spontaneous speech, and only in 13% of full stops and 42% of commas for read speech. Among all occurrences of filled pauses, 8% indicate full stops and 6% indicate commas, among breath pauses the proportions are, respectively, 10 and 11% (Igras, Ziółko, 2013a; 2013c).

4.3. Energy and power

Speakers usually tend to quiet loudness of their voice at the end of the sentence. The parameters describing energy changes were most often used in sentence boundary detection, although some authors do not consider them (due to their dependence on channel variability: Shriberg *et al.*, 2000). We measured the phenomenon on the phoneme level (Igras, Ziółko, 2014b). Using statistical values of energy (root mean square, RMS) and power (amount of energy per time unit) of each phoneme class realizations, we elaborated statistical models of probability that the phoneme with given relative energy and power, appears in the end of sentence (Fig. 3). We found that the average relative value of energy of the last phoneme in a sentence was

a)



b)

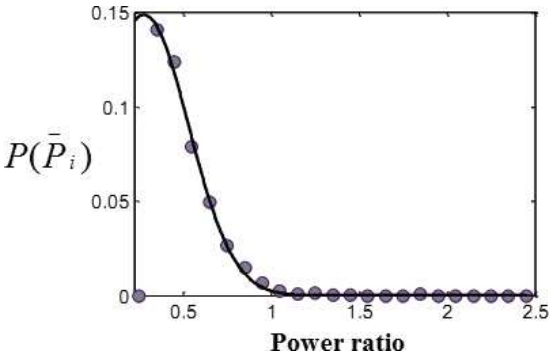


Fig. 3. Probability of the phoneme with a given energy (a) or power ratio (b) to be the last in the sentence.

0.6 of the average value of the phoneme class, while the relative power – 0.36. Therefore, the following parameters are computed (in the parentheses: abbreviation and number of the parameter are included):

- frame energy (RMS), (*en*, 1);
- ratio of the frame energy to the mean frame energy in the phrase (*en_ratio*, 2);
- place of the frame in the ranking of the most silent frames in the phrase (*en_rank*, 3);
- mean of relative energy values of the phonemes in the frame (*en_phon_ratio*, 4) or of the two last phonemes in the frame (*en_phon_ratio*, 5);
- mean of probability values based on energy of all phonemes in the frame (*en_prob*, 6) or of two last phonemes (*en_prob_last*, 7);
- slope of the linear regression of the relative values of the phonemes in the frame (*en_ratio_slope*, 8);
- standard deviation of the 25 ms sub-frames energy within the frame (*en_fr_std*, 9) and slope of their linear regression (*en_fr_std*, 10);
- mean phoneme value in the frame (*pow_phon_mean*, 11) or mean of two last phonemes (*pow_phon_last*, 12);
- mean relative power of all phonemes in the frame (*pow_phon_ratio_mean*, 13) or mean of two last phonemes (*pow_phon_ratio_last*, 14);
- mean of probability values based on power of all phonemes in the frame (*pow_prob*, 15) or of two last phonemes (*pow_prob_last*, 16);
- slope of the linear regression of the relative power values of the phonemes in the frame, (*pow_ratio_slope*, 17);
- slope of the linear regression of the probability values based on power of the phonemes in the frame (*pow_prob_slope*, 18);
- slope of linear regression of power values of the 25 ms sub-frames within a frame (*pow_fr_slope*, 19);
- how many of 10 phonemes of the lowest power in the phrase the frame contains (*pow_min_nr*, 20).

4.4. Phonemes length

The regularity of lengthening last phonemes in the sentence was reported for many languages. Therefore, the duration features were always included in sentence boundary detection. We examined the phenomenon statistically taking into account also differences between speakers (Igras *et al.*, 2013b; 2014a). In one of the experiments of boundary detection based only on phoneme duration features, we obtained correct detection of 37% of sentence ends, with only around 2.5% rate of false detections (Igras *et al.*, 2013b). The final model is presented in Fig. 4. The following features were extracted:

- mean duration of all phonemes in the frame (*dur_mean*, 21) or of the two last phonemes (*dur_last*, 22);
- relative duration value of the last vowel in the phrase (*dur_last_vow*, 23);
- mean relative duration of all phonemes in the frame (*dur_phon_ratio*, 24) or of the last two phonemes (*dur_phon_ratio_last*, 25);
- mean of probability values based on duration for all phonemes in the frame (*dur_prob*, 26) or for the last two phonemes (*dur_prob_last*, 27);
- slope of linear regression of probabilities based on duration for all phonemes in the frame (*dur_prob_slope*, 28);

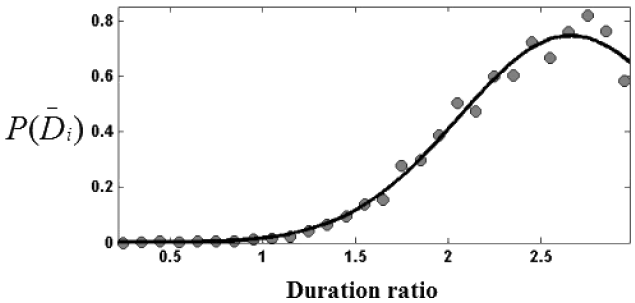


Fig. 4. Probability of the phoneme with given duration ratio to be the last in the sentence.

- how many of 10 phonemes of the highest relative duration in the phrase the frame contains (*dur_rank*, 29);
- ratio of the mean relative duration in the frame to the mean for the phrase (*dur_mean_ratio*, 30).

4.5. Speech rate

Inversely to the phone duration, speech rate should slow down towards the end of the sentence. We chose the following parameters to describe speech rate changes:

- number of phonemes in the frame (*sr_phon*, 31);
- number of vowels in the frame (*sr_syl*, 32);
- ratio of number of phonemes (*sr_phon_ratio*, 33) or vowels (*sr_syl_ratio*, 34) in the frame to the mean for the phrase;
- place of the frame in the ranking of the slowest speech rate frames in the phrase, (*sr_phon_rank*, 35).

4.6. Fundamental frequency

Falling intonation is one of the most easily perceived markers of sentence boundary. Physiologically, at the end of phrase the decreasing glottal pressure causes cadence of pitch contour. We examined changes of F0 contours in the ends of affirmative sentences (Igras, Ziółko, 2014b). A mean decrease of the F0 within a sentence from CORPORA was about 98 Hz (slope of linear regression −0.36). F0 was decreasing for 97% of sentences.

For pitch tracking, YAAPT algorithm was used (Zahorian, Hu, 2008) and for each frame we computed:

- ratio of mean pitch value in the frame to mean pitch in the phrase (*F0_mean_rel*, 36);
- maximal (*F0_max_rel*, 37), minimal (*F0_min_rel*, 38) or last (*F0_end_rel*, 39) pitch value divided by the mean of the phrase;
- difference between first and last value of pitch (*F0_start_end*, 40);
- difference between maximal and minimal pitch value in the frame (*F0_amp*, 41);
- slope of linear regression of pitch within the frame (*F0_slope*, 42) or within the second half of the frame (*F0_slope*, 43);
- slope of linear regression of pitch in the second half of the frame divided by the slope of the whole phrase (*F0_slope_last_rel*, 44);
- mean pitch of the frame divided by the mean of the first frame in the phrase (*F0_slope_last_rel*, 45).

4.7. Frequencies in subbands

According to perceptual observation, a sentence beginning is often characterized by higher frequencies than the rest of the utterance. To describe this tendency, discrete wavelet transform (DWT) with mel-scale was used (Ziółko *et al.*, 2011). Extracted parameters describe amount of energy in 12 subbands obtained after decomposition:

- energy in the frequency subbands 1–12 (*subb_en_1*–*subb_en_12*, 46-57);
- energy in the frequency subbands 1–12 in the second half of the frame (*subb_en_last_1*–*subb_en_last_12*, 58–69);
- difference in amount of energy between first and second half of the frame (*subb_en_diff_1*–*subb_en_diff_12*, 70–81);
- sum of products of multiplication of the subband number (1–12) and place in the ranking of the subbands energy, in descending order (*subb_rank_score*, 82);
- difference in *subb_rank_score* between first and second half of the frame (*subb_rank_score_diff*, 83).

4.8. Mel-Frequency Cepstral Coefficients (MFCC)

Majority of ASR systems work on the basis of MFCC, so we verified their usability for punctuation detection. MFCC were computed for 20 ms frames (with 1/2 shift), using 13 filterbank channels and the frequency range 75–4000 Hz. The 0th coefficient was not included. The following parameters were extracted:

- MFCC values (12 coefficients + deltas) for the frame (*mfcc_1*–*mfcc_24*, 84–107);
- differences in MFCC values between first and second half of the frame – 12 values + deltas (*mfcc_rel_1*–*mfcc_rel_24*, 108–131).

5. Factor analysis

Many of 132 features listed in previous section are correlated with each other, what makes them redundant in describing acoustic properties of sentence ends. After analysis of their covariance, we performed Principal Component Analysis (PCA). First, we normalized features using Z-standardization to obtain normal distribution of mean = 0 and standard deviation = 1.

The same number of observation frames (5130) from each class (the first frame, the last frame and the random frame that is not located in the beginning or in the end) were taken from the training set. PCA allowed reducing dimensionality from 132 to 20 principal components, which explain over 96% of variation in the data set. The coefficients of the components are presented graphically in Fig. 5.

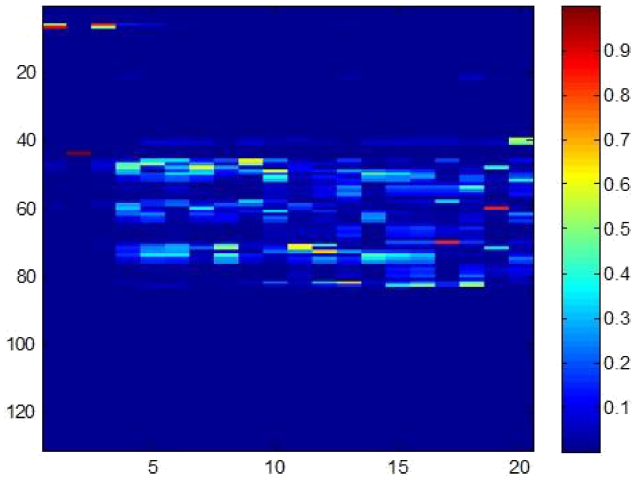


Fig. 5. 20 principal components of 132 features.

6. Evaluation metrics

The choice of an appropriate evaluation measure was often put into question. Although recall and precision are standard informative measures of classification performance, they were criticized for underweighting missing and spurious identification errors compared with incorrect identification errors (Makhoul *et al.*, 1999). The new measure, slot error rate (SER), was proposed. It equally weighted three types of identification error (incorrect, missing, and spurious) and was applied for sentence boundary evaluation (Gotoh, Renals, 2000). SER is analogical to WER. It is obtained by formula

$$\mathrm{SER} = \frac{(I+M+S)}{(C+I+M)}, \tag{1}$$

where C, I, M, and S denote the numbers of correct, incorrect, missing, and spurious identifications. Using this notation, recall and precision scores may be calculated as

$$\begin{aligned} P &= \frac{C}{(C+I+M)}, \\ R &= \frac{C}{(C+I+S)}, \end{aligned} \tag{2}$$

respectively. The lower the SER score, the better. Another measure used for this task was boundary classification error, i.e. the percentage of word boundaries labeled with the incorrect class (Shriberg *et al.*, 2000). The next scoring tool is SU error rate (number of incorrect boundaries divided by the total number of SU boundaries).

For the needs of our work (we recognize only absence or presence of punctuation mark, with no distinction to classes), we adopted the classical measures of precision and recall

$$\begin{aligned} P &= \frac{C}{(C+M)}, \\ R &= \frac{C}{(C+S)}, \end{aligned} \tag{3}$$

and F-measure

$$F = \frac{2PR}{(P+R)}. \tag{4}$$

7. Classification

Several classifiers were tried and compared in the task of frames classification into one of 3 classes: <start>, <fstop> or <non>:

- Classification-Regression Decision Trees (CART), as frequently used in previous works, nonparametric classifier based on simple rules. Initially built decision tree was pruned (at the level = 34);
- multi-class Discriminant Analysis (Linear – LDA and Quadratic – QDA);
- Naive Bayes classifier (NB), using Gaussian models of features distribution;
- k-nearest Neighbors classifier (kNN) as one of the simplest machine learning classifier. The tests showed that the best results were obtained for $k = 1$ and the city block distance measure.

Table 2 shows performance of these classifiers measured with resubstitution error and cross-validation error. The first row contains the resubstitution error, which is the proportion of misclassified observations on the training set. In the second row we included a stratified 10-fold cross-validation error. For calculating the cross-validation error, we randomly divided the training set into 10 disjoint test subsets (each subset has roughly equal size and roughly the same class proportions as in the training set).

Table 2. Comparison of classifiers using resubstitution error (r. error) and cross-validation error (c.-v.error).

Measure	CART	LDA	QDA	NB	kNN
r. error	6.63	37.24	32.31	**3.27**	0.65
c.-v. error	9.40	37.44	33.33	**3.33**	25.49

The best results were obtained for Naive Bayes classifier. It achieved 86% sensitivity and 95% specificity.

Then, in the experiment of automatic boundaries detection we performed evaluation of the previously trained classifiers. The testing set was used, but we did not distinguish classes of punctuation marks. Commas and full stops were treated jointly as an appearance of boundary. The results are grouped in Table 3. Although the best recognition level was achieved for

Table 3. Comparison of classifiers' performance on the testing set.

Speaker	Measure	CART	LDA	QDA	NB	kNN
1	Precision	25.15	38.10	24,29	**43.82**	24.55
	Recall	97.62	19.05	80.95	**92.86**	97.62
	F-measure	40.00	25.40	37.36	**59.54**	39.23
2	Precision	40.00	33.33	37.61	**72.16**	38.95
	Recall	100.00	1.43	58.57	**100.00**	95.71
	F-measure	57.14	2.74	45.81	**83.83**	55.37
3	Precision	16.88	20.00	16.92	**40.40**	17.18
	Recall	100.00	2.50	82.50	**100.00**	97.50
	F-measure	28.88	4.44	28.09	**57.55**	30.12
All	Precision	26.00	32.14	23.98	**51.94**	26.08
	Recall	99.33	5.00	70.67	**98.00**	96.67
	F-measure	41.22	10.11	35.81	**67.90**	41.08

Naive Bayes classifier, all the classifier showed tendency to high recall rate and relatively low precision.

8. Discussion

In our approach we combined inductive and analytical learning. Some *a priori* knowledge was available in the form of rules known from phonetic and phonological literature (e.g. preboundary lengthening or quietening voice in the end of sentence). It helped for feature selection and modeling numerically the phenomena. Using a set of training observations, some other regularities were found empirically. Fusion of the two sources of information allowed to generalize the relationship between acoustic characteristics of speech and language syntax (punctuation marks) and finally to build classifiers.

We adopted direct prosody modeling approach (Shriberg, Stolcke, 2004), which did not require any hand-labeling of prosodic structures. Observation window in this work is 0.5 s, while other authors utilize features on word level or, e.g. 200 ms before and after inter-word boundary. We use feature of the frame itself as well as features computed in reference to the whole phrase or neighboring frames. Besides standard features, some new ones are suggested.

The obtained results seem to be acceptable, especially taking into account nature of Polish punctuation, discussed in Subsec. 2.2. The weakness of our method is relatively low precision ratio, caused by too many insertions (false positives). The possible reason was the style of speech in testing set (careful pronunciation that indicated clear phrasing not only in places of punctuation marks). Nevertheless, the insertions usually appeared in the end of speech phrases that formed semantically consistent entities. It would seem to confirm that the natural unit seems to be the phrase, rather than the sentence (Gotoh, Renals, 2000). Nevertheless, it was also proved by (Cole *et al.*, 2010) that there is a close relationship between perceived prosodic boundaries and syntactic structure, and this relationship is mediated partly by acoustic encoding of prosody.

A weakness of the study might have been also treating together punctuation marks (full stops and commas). There may be some significant differences in the acoustic realization of the boundaries of these two phrasing levels. It could have a negative impact on the achieved results.

Modeling based on prosody reveals several problems. It was pointed (Kolar, Lamel, 2011) that there is a trade-off relation between prosodic means: a weaker use of one prosodic mean can be compensated by a stronger use of another. Prosody is affected by emotions or a speaker individual style. In natural speech, phrases structures differ in length or intonational patterns, due to sentence syntax, number and type of accents.

9. Conclusion

We investigated acoustic correlates of punctuation in spoken Polish in order to automatize punctuation insertion into ASR transcripts. We analyzed groups of parameters describing changes of pitch, energy, power, duration and subband frequency in the end of sentences, as well as some pauses and phrase duration statistics. For the most significant parameters, probabilistic models were built in order to describe tendencies known from phonetic and phonological literature. Performance of several machine learning methods was compared in the task of automatic detection of sentence ends, using annotated corpus of read speech. The most efficient, in the terms of f-measure, was Naive Bayes classifier with score of 67.90%. The high recall and decent precision leave good hopes on additional (eg. NLP) methods can improve this method even further.

The study was conducted as a pioneer one for Polish. Its preliminary results are satisfactory, but it still requires a further development. The models and classifiers need to be verified using more demanding spontaneous speech, characterized by irregular prosody. In future works more attention will be paid to commas. We plan to build separate classifiers for sentence boundaries ended with a full stop and clause boundaries ended with a comma. An additional study must be conducted in the area of punctuation detection from the textual cues, and finally the fusion of the acoustic model and language model will result in more complex approach to the task. Optimization of the algorithm and integration with the current ASR system is also planned, as well as releasing the evaluation data for other scientists who wish to study this problem.

Acknowledgments

The project was funded by the National Science Centre allocated on the basis of a decision DEC-2011/03/D/ST6/00914. We would like to thank J. Grzybowska and S. Kacprzak for their comments and suggestions, which helped us prepare this paper.

References

1. Barczewska K., Igras M. (2013), *Detection of disfluencies in speech signal*, Challenges of Modern Technology, **32**, 1–2, 127–154.
2. Baron D., Shriberg E., Stolcke A. (2002), *Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues*, Proceedings of the International Conference on Spoken Language Processing, 949–952.
3. Batista F., Caseiro D., Mamede N., Trancoso I. (2008), *Recovering capitalization and punctuation marks for automatic speech recognition: Case study for Portuguese broadcast news*, Speech Commun., **50**, 10, 847–862.
4. Beeferman A.B.D., Lafferty J. (1998), *Cyberpunc: a lightweight punctuation annotation system for speech*, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 689–692.
5. Chistikov P., Khomitsevich O. (2013), *Improving prosodic break detection in a Russian TTS system*, Speech and Computer, ser. Lecture Notes in Computer Science **8113**, Springer International Publishing, 181–188.
6. Christensen H., Gotoh Y., Renals S. (2001), *Punctuation annotation using statistical prosody models*, Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding, 35–40.
7. Cole J., Mo Y., Baek S. (2010), *The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech*, Language and Cognitive Processes, **25**, 7–9, 1141–1177.
8. Demenko G., Wagner A. (2007), *Prosody annotation for unit selection TTS synthesis*, Archives of Acoustics, **32**, 1, 25–40.
9. Demenko G. (1999), *Analysis of Polish Suprasegmentals for Speech Technology* [in Polish: *Analiza cech suprasegmentalnych jezyka polskiego na potrzeby technologii mowy*], ser. Seria Jezykoznawstwo Stosowane, Poznań, Wyd. Naukowe UAM.
10. Dłuska M. (1976), *Polish prosody* [in Polish: *Prozodia języka polskiego*], Warszawa, Państwowe Wydawnictwo Naukowe.
11. Fach M.L. (1999), *A comparison between syntactic and prosodic phrasing*, Proceedings of the European Conference on Speech Communication and Technology, 527–530.
12. Frąckowiak-Richter L. (1973), *The duration of Polish vowels. Speech analysis and Synthesis III*, PWN.
13. Gotoh Y., Renals S. (2000), *Sentence Boundary Detection in Broadcast Speech Transcripts*, Proc. of ISCA Workshop: Automatic Speech Recognition: Challenges for the new Millennium, 228–235.
14. Grabe E., Karpiński M. (2003), Universal *and language-specific aspects of intonation in English and Polish*, Proceedings of the 15th International Congress of Phonetic Sciences, 3–9 August, Barcelona, 1061–1064.
15. Grocholewski S. (1997), *CORPORA – speech database for Polish diphones*, Proceedings of Eurospeech.
16. Huang J., Zweig G. (2002), *Maximum Entropy Model for Punctuation Annotation from Speech*, Proc. International Conference on Spoken Language Processing (ICSLP).
17. Igras M., Ziółko B., Jadczyk T. (2012), *Audiovisual database of Polish speech recordings*, Studia Informatica, **33**, 2B, 163–172.
18. Igras M., Ziółko B. (2013a), *Different types of pauses as a source of information for biometry*, MAVEBA, Florence, 197–200.
19. Igras M., Ziółko B., Ziółko M. (2013b), *Length of phonemes in a context of their positions in polish sentences*, Proceedings of SIGMAP, the International Conference on Signal Processing and Multimedia Applications, 59–64.
20. Igras M., Ziółko B. (2013c), *Wavelet method for breath detection in audio signals*, Multimedia and Expo (ICME), IEEE International Conference on.
21. Igras M., Ziółko B., Ziółko M. (2014a), *Is phoneme length and phoneme energy useful in automatic speaker recognition?*, XXII Annual Pacific Voice Conference, Krakow.
22. Igras M., Ziółko B. (2014b), *Role of acoustic features in marking stress and delimiting sentence boundaries in spoken Polish*, Acta Physica Polonica A, **126**, 6, 1246–1257.
23. Jassem W. (1962), *Accent of Polish* [in Polish: *Akcent języka polskiego*], Wrocław, Ossolineum.
24. Jassem W. (1973), *Rudiments of acoustic phonetics* [in Polish: *Podstawy fonetyki akustycznej*], Warszawa, Państwowe Wydawnictwo Naukowe.
25. Karpowicz T. (2012), *Culture of Polish: pronunciation, ortography, punctuation* [in Polish: *Kultura języka polskiego: Wymowa, ortografia, interpunkcja*], Wydawnictwo Naukowe PWN, Warszawa.
26. Kim J.H., Woodland P.C. (2003), *A combined punctuation generation and speech recognition system and its performance enhancement using prosody*, Speech Communication, **41**, 4, 563–577.
27. Klessa K., Karpiński M., Kleśta J. (2002), *A preliminary study of the intonational phrase, nuclear melody and pauses in Polish semi-spontaneous narration*, Speech Prosody Proceedings.
28. Klessa K., Śledziński D. (2008), *A study of chosen temporal relations within syllable structure in Polish*, Speech and Language Technology.

29. Klessa K. (2011), *Polish segmental duration: selected observations based on corpus data*, Speech and Language Technology, Special Issue dedicated to Wiktor Jassem, 94–104.

30. Kolar J., Svec J, Psutka J. (2004), *Automatic punctuation annotation in Czech broadcast news speech*, Saint-Petersburg, SPIIRAS, 319–325.

31. Lööf J., Gollan C., Ney H. (2009), *Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a Polish speech recognition system*, Proceedings of Interspeech, Brighton, 88–91.

32. Łuczyński E. (1999), *Contemporary Polish punctuation: a norm and an usus* [in Polish: *Współczesna interpunkcja polska. Norma a uzus*], Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk.

33. Makhoul J., Kubala F., Schwartz R., Weischedel R. (1999), *Performance measures for information extraction*, Proceedings of DARPA Broadcast News Workshop, pp. 249–252.

34. Malisz Z., Wagner P. (2012), *Acoustic-phonetic realisation of Polish syllable prominence: a corpus study*, Rhythm, melody and harmony in speech. Studies in honour of Wiktor Jassem, 105–114.

35. Navas E., Hernez I., Sainz I. (2008), *Evaluation of automatic break insertion for an agglutinative and inflected language*, Speech Communication, **50**, 1112, 888–899.

36. Ostaszewska D., Tambor J. (2000), *Phonetics and phonology of modern Polish language* [in Polish: *Fonetyka i fonologia współczesnego języka polskiego*], PWN.

37. Przyłubski F. (1953), *A few words on the history of comma* [in Polish: *Kilka słów o historii przecinka*], Poradnik Językowy, **8**.

38. PWN Dictionary 2013, *Polish spelling and punctuation rules* [Online], http://so.pwn.pl/zasady.php?id=629737 Accessed: 19/05/2015.

39. Shriberg E., Stolcke A., Hakkani-Tür D., Tür G. (2000), *Prosody-based automatic segmentation of speech into sentences and topics*, Speech Communication, **32**, 1–2, 127–154, 10.1016/S0167-6393(00)00028-5.

40. Shriberg E., Stolcke A. (2004), *Direct Modeling of Prosody: An Overview of Applications in Automatic Speech Processing*, Proceedings of International Conference on Speech Prosody, Nara, Japan.

41. Steffen Batóg M. (1996), *Structure of melody contour in Polish* [in Polish: *Struktura przebiegu melodii polskiego języka ogólnego*], Poznan, Wydawnictwo UAM.

42. Stevenson M., Gaizauskas R. (2000), *Experiments on Sentence Boundary Detection*, Proc. Conference on Applied Natural Language Processing (ANLP), 84–89.

43. Wang D., Lu L., Zhang H.J. (2003), *Speech segmentation without speech recognition*, Acoustics, Speech, and Signal Processing Proceedings (ICASSP '03). 2003 IEEE International Conference on, 468–471.

44. Wang D., Narayanan S.S. (2004), *A multi-pass linear fold algorithm for sentence boundary detection using prosodic cues*, ICASSP.

45. Vicsi K., Szaszak G. (2006), *Prosodic cues for automatic phrase boundary detection in ASR*, Proceedings of the 9th International Conference on Text, Speech and Dialogue, ser. TSD'06, Berlin, Heidelberg: Springer-Verlag, 547–554.

46. Wagner A. (2008), *A comprehensive model of intonation for application in speech synthesis*, Dissertation, Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza, Poznań.

47. Wagner A. (2010), *Acoustic cues for automatic determination of phrasing*, Proceedings of Speech Prosody.

48. Wagner A., Bachan J., Klessa, K., Demenko G. (2015), *Przegląd wybranych aspektów analizy prozodii mowy spontanicznej na potrzeby technologii mowy* Prace Filologiczne, (LXVI), 271–298.

49. Wypych M. (2011), *A system recognizing intonation structures in speech signal* [in Polish: *Układ rozpoznający struktury intonacyjne w sygnale mowy*], Dissertation, PAN, Warszawa.

50. Zahorian S.A., Hu H. (2008), *A spectral/temporal method for robust fundamental frequency tracking*, The Journal of the Acoustical Society of America, **123**, 4559–4571.

51. Ziółko M., Gałka J., Ziółko B., Jadczyk T., Skurzok D., Mąsior M. (2011), *Automatic speech recognition system dedicated for Polish*, Proceedings of Interspeech, Florence.

52. Ziółko B., Ziółko M. (2011), *Time durations of phonemes in Polish language for speech and speaker recognition*, Human Language Technology. Challenges for Computer Science and Linguistics. Lecture Notes in Computer Science, 6562/2011, 105–114.