The Role of Acoustic Features in Marking Accent and Delimiting Sentence Boundaries in Spoken Polish

M. Igras and B. Ziółko
Faculty of Computer Science, Electronics and Telecommunications, Department of Electronics,
AGH University of Science and Technology, al. A. Mickiewicza 30, 30-059 Krakow, Poland

(Received November 28, 2013; revised version August 14, 2014; in final form October 22, 2014)

In this article the authors investigated and presented statistical models of acoustic phenomena observed within realizations of phonemes and the correlations of the acoustic properties with functional features, such as accents and sentence boundaries. The authors used two databases: the first one contained separately produced sentences and the second one — phrases extracted from larger, continuous stretches of natural speech. The authors also statistically analyzed the selected features of Polish phonemes' realizations (the duration, energy and power of the phonemes, the fundamental frequency of voiced phones) in order to detect their relations with the phone location in a sentence. Additionally, the authors built the probabilistic models and suggested the evaluation methods to assess quantitatively the phenomena known from phonetic literature. Finally, the authors have identified the pre-boundary lengthening of the phones and a decrease of energy and pitch as the markers of sentence endings. In the place of accented syllables, we have observed a significant increase of total energy and power, accompanied by a local increase of F0. Finally, we have indicated possible application of the results for speech technology.

DOI: 10.12693/APhysPolA.126.1246
PACS: 43.72.+q, 43.55.Ka, 43.72.Ar, 43.72.Ne

1. Introduction

In the automatic speech recognition (ASR) system, models of particular phonemes are created to build a database of patterns. Typically, the models are based on low-level features, e.g. mel-frequency cepstral coefficients (MFCC) [1] or wavelet coefficients [2]. High-level features, like prosody-related ones, are rarely included. All the acoustic features of realizations of the same phoneme can differ for many reasons. The most important are speaker-connected factors (a given speaker's individual features [3], their emotional state, dialect [4] and potential speech pathology [5]) and language-connected factors (neighborhood of other phonemes [6], syllable structure [7], information and discourse structure [8], as well as the accentuation and the phone position in a sentence). In this work, we focus on the two latter factors.

The goal of the study is to measure and model how prosodic characteristics of phonemes' realizations differ depending on their position in a sentence (in the ending of a phrase and under lexical stress). The main application of the research is language modeling for the Polish ASR system. It was already shown by Demenko et al. [1] that including the stressed vowel model in the dictionary level (for each vowel a model of a stressed and a model of an unstressed equivalent was applied) into the ASR reduces the word error rate (WER). Also, by studying changes in phoneme's realization in sentence endings, we should be able to develop algorithms for the automatic punctuation detection in spoken language. Improving the ASR with the automatic insertion of punctuation marks makes voice interface more comfortable (without the need for dictating punctuation marks), increases transcripts' legibility and enables us to adapt them directly for natural language processing systems. Hence, both the prosodic features and their correlation with position in a sentence are speaker-specific, which makes them useful for speaker biometry systems. The obtained models can be also used to make synthesized speech sound more naturally.

In this paper, we briefly summarize description of the investigated phenomena from phonetic and phonological literature (Sects. 2, 3), then we introduce our approach (Sects. 4, 5), describe obtained numerical models (Sects. 6, 7) and discuss them in comparison to other results (Sect. 8).

2. Correlates of punctuation in spoken language

Polish, as a Slavic language, is characterized by a relatively free word order and a complex inflectional morphology. Polish punctuation system is described as mainly syntactic, though as an additional factor the influence of rhythmic structure and intonation of speech are also included (contemporary dictionaries like [9] or handbooks of punctuation like [10]). Likewise, authors of recent elaborations on modern Polish emphasize logical-structural role of punctuation, and as the secondary aspect the prosody (intonation, pauses, breathing) is taken into account [11]. Therefore, the role of the proper punctuation usage covers: exact expression of author intention, suggesting intonation for a reader, indicating places where pause should occur [10].

Historically, Polish writing system was borrowed from Medieval Latin. Similarly to Latin, in the earliest Polish writings points were used to mark breaks: the full stop for a long pause, the colon for a medium one and the comma for the shortest one, in order to facilitate the usage of pauses for the reader [12]. Punctuation rules were codified in XVIII century focusing on the semantic structure, but also on the rhythmic structure of speech or on

the emotional tone. As it evolved, our language seemed to lose the initially close connection with prosody in favor of mirroring the semantic and syntactic structure. Therefore, 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 [13].

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. For other languages, acoustic correlates of phrase endings and sentence endings were investigated in detail in order to build systems that automatically insert punctuation into transcripts (punctuation annotation systems) [14–16].

Acoustic events occurring at the end of a phrase are related with the physiological background of speech production. Dynamic breathing determines the capability of uttering full sentences or phrases within a breath [17]. As a result, realizations of the same phoneme, localized in different places of a sentence, differ in acoustic quality. At the end of the breathing phase, articulation is usually weakened (less precise, quieter, with lower vocal effort).

For Polish, a study of intonation phrases' boundaries was conducted by Wagner [18, 19] and Demenko and Wagner [20], 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% [19].

It was observed that speakers usually tend to slow down their speech towards the endings of utterance units. Speaking rate depends on individual characteristics of a person, but typically the last phones in a sentence are much longer. Also, for the automatic punctuation annotation in Czech [15] the phenomenon of pre-boundary lengthening is applied (only for vowels). It is described by the following parameters: average duration of vowels, duration of the first and the last vowel and duration of the longest and the shortest vowel. The phenomenon was described for Hungarian [21] and Japanese [22] as well.

Prosody has been a subject of study of many phoneticians. Some recent works on Polish prosody were provided by Klessa et al. [4, 7, 8], Karpiński et al. [8, 23], Wagner [18–20], Malisz [24], Wypych [25] and Demenko [26]. In case of the latter, not only statistical analysis was introduced, but also automatic classification of intonational phrase structure as well as accentual structures was performed. Models of suprasegmental structures were applied to the speech synthesis. The detailed analysis of the correlations between the place in a phrase and the phonemes' duration influenced by accent and intonational context was also described in [26]. The research showed that, in Polish, the last and the one before the last vowel in a sentence are lengthened regardless of accent presence, but the lengthening effect is stronger for accented syllables. Phrase perception depending on phoneme duration was studied on logatomes [27]. The dependence of segmental duration on a selection of interacting factors is discussed by Klessa [4]. In another paper she finds out that the duration of particular syllable parts depends significantly on the presence of word stress and the syllable's position with respect to pauses [7]. Most of the cited works investigate phenomena connected to intonation phrases, but there is a lack of a complex study that would define prosodic indicators of punctuation marks.

For further discussion it is important to declare that in our research data from corpus I (further detailed in Sect. 4), we process utterances short enough to be assumed as acoustic phrases. For corpus II, we consider as phrases sections of speech separated in their transcripts by commas or full stops. This approach is justified by the aim of our research: looking for prosodic correlates of punctuation marks in spoken language. Thanks to this assumption we will be able to observe how acoustic events in speech signal correlate with commas and full stops in speech transcripts.

3. Accents in Polish

Polish accentual system has a regular word stress usually on the penultimate syllable of the word. There is no fixed sentence accent. The accent's role is mainly semantic and syntactic [28, 29].

Nature of accentuation in Polish has not been clearly defined. Some sources characterize it as pitch and dynamic accent, where the term "pitch" describes changes of auditory correlate of voice fundamental frequency and the word "dynamic" describes changes of voice intensity as means of expression [30]. Additionally, a rhythmic type of accent (based on the change of the phone duration) was described for Polish [30, 31]. Jassem [29] investigated four acoustic features in terms of indicating the place of accent: intensity, pitch, duration and voice timbre. He found that only intonation (pitch contour) shows regularity connected with accented syllables and he classified Polish accent as the pitch accent. The significance of acoustic parameters in perception of accents was also investigated by Demenko [26]. The variance analysis confirmed that the following features are meaningful: an interval of fundamental frequency changes within a vowel, a relative change of fundamental frequency value in comparison to the amplitude of F0 within the whole phrase, the F0 relation to speaker specific F0 values, vowel durations (in case of the postictic accent). As a result of the experiment it was concluded that accents can be recognized only with acoustic cues, without taking into account contextual cues. The F0 features were proved to be crucial in the main accent detection, while duration features are important for subsidiary accents' signalization. Malisz and Wagner [24] suggest that overall intensity, duration and pitch movement are good correlates of phrase accent as well. The influence of stress

on the segmental duration was confirmed in [32] (in accordance with the works by Jassem [29] and Dłuska [33]). Three duration classes were distinguished: the stressed vowel (the longest), the word final vowel (comparably shorter), and the vowel in a pre-stressed syllable or a post-stressed syllable but not word-final (the shortest). On the other hand, in Klessa et al. investigation based on the Polish Intonational Database PoInt, quite a random distribution of duration values was demonstrated. It did not seem to confirm any participation of duration in the expression of word stress in Polish [8].

Stress in Polish is regarded as "weakly expressed" acoustically, when compared with other Slavic languages [34]. It was supported by other results: having analyzed spontaneous speech from task-oriented dialogues corpus, Malisz and Wagner confirmed that the fixed penultimate stress pattern in Polish is determined linguistically as a highly influential "expectation" that is perceived by native listeners but not attested acoustically in a clear way. They indicate that intensity difference and maximum pitch difference are the main determinants of overall prominence. They confirm weak stress effects on duration [24].

Intonation patterns in Polish declarative sentences are rather flat, monotonous and a regular fall is noticeable on the last prominent word. In a contrastive study of Polish and British English, six main tunes (complete pitch patterns of sense groups) are distinguished. Among them, medium rise and medium fall is typical for Polish and high rise and high fall is characteristic for English and the rest are present in both, English and Polish (low rise, fall-rise, rise-fall, low fall). Even for the common tunes, the pitch amplitude of an average Pole is not as high as for British [35]. For the enumeration, the intonation patterns are similar. Another cross-linguistic study [23] pointed that Polish speakers produce a smaller range of nuclear accent types than English speakers and widest range of contour types in declaratives (predominantly LM type). 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.

A detailed perceptual study on Polish accentual structures was delivered by Steffen-Batogowa [36]. It was pointed that the number of main accents in an utterance depends on quality of used words: their parts of speech type, their length (measured by number of syllables), syntactic and semantic relations of neighboring words. Among additional factors that influence accentual structure, there are speech rate, presence of strong logical accent, style of spoken language or individual style of speakers.

Different types of pitch accent in Polish were modeled for purposes of speech synthesis [37] using classification and regression trees (CART). The accuracy of accent prediction was 83.3%.

Lately it was showed that Polish lacks rhythmic secondary stress [38]. In our paper, we focus on a main accent in orthotonic words, not taking into account clitics or a secondary accent in an orthotonic word.

4. Data

In order to collect statistics on the varied acoustic realizations of sentence ends and accents, we used two corpora. The first one contained isolated (produced as separate units) sentences, while the second sentences were extracted from larger, continuous stretches of speech. The decision was influenced by the relatively high number of different speakers and the availability of phoneme annotations.

The first corpus (Db1) was a part of Polish speech database CORPORA [39] consisting of isolated sentences of 45 speakers (male, female, and children). The content of utterances was designed to obtain the best possible Polish diphones distribution resulting in semantically artificial character of the sentences. In total, the Database 1 contains 5130 sentences/phrases (98111 phones). The data were recorded with condenser and dynamic microphones, in standard room conditions, in the presence of a working computer.

As the second corpus (Db2), we prepared fifteen minutes of speech (3 speakers), which were annotated manually on phoneme level, and punctuation marks were included in the annotation. Recordings from AGH Audio-Visual Speech Database [40] and one recording of a monologue were used. All the utterances were informative speech: read or spoken after preparation. The Database 2 contains 116 sentences/ 254 phrases (9797 phones). Similarly, the recordings were made in a relatively standard office room with dynamic or condenser microphones using the ZOOM H4n recorder.

We also used time annotations of the recordings. Each recording has an mlf file (master label file, HTK 3.0 standard) attached. It contains information on time of the beginning and the end of each phoneme realization (an example is presented in Table I). Some of the annotations were made manually while the majority of them by an automated computer program.

For both databases, Db1 and Db2, we used the CORPORA notation of phonemes (its correspondence to SAMPA and IPA standards was prepared on the basis of [28] and collected in Table II). It distinguishes 37 phoneme classes, with no distinction of positional variants (like those described in [28] or [41]).

5. Methods

First, we gathered acoustic features' statistical values from CORPORA using 114 sentences, each produced by 45 speakers. Each recorded sentence was obtained as a discrete signal $s(n)$ describing changes of acoustic pressure in time, where n is the order number of a given sample (the sampling frequency f_s was 16 kHz). For each phonetic segment i, which is the realization of m-th of 38 classes of phonemes, we calculated the data on:

TABLE I

An example of an mlf with annotation.

"*/ak1c1001.lab"		
0	100000	sil
150000	1200000	l
1250000	1800000	u
1850000	2450000	b
2500000	3150000	i
3200000	3650000	ci
3700000	4450000	cz
4500000	5300000	a
5350000	6050000	r
6100000	7250000	d
7300000	7750000	a
7800000	8850000	sz
8900000	9900000	o
9950000	10450000	w
10500000	11250000	y
11300000	12200000	p
12250000	12700000	l
12750000	15550000	a_
15600000	17700000	s
17750000	18050000	sil

— segment duration, calculated on the basis of annotation files

$$D_{m,i} = \frac{n_{\text{end}} - n_{\text{start}}}{f_{\text{s}}}, \tag{5.1}$$

— segment energy, as a square root of a sum of samples' values (RMS):

$$E_{m,i} = \sqrt{\sum_{n=n_{\text{start}}}^{n_{\text{end}}} s^2(n)}, \tag{5.2}$$

— segment power, dividing energy of a segment by its time duration:

$$P_{m,i} = \frac{E_{m,i}}{D_{m,i}}. \tag{5.3}$$

These features are a classical measure used by a vast number of researchers in signal processing field [42–43].

After calculating the three features for each phone, we verified distribution of their values within phoneme classes (some examples of histograms are presented in Table III).

We discovered that the distribution of these features can be modeled with lognormal distribution

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp - \left(\frac{\ln x - \mu}{2\sigma^2}\right)^2, \tag{5.4}$$

where x is the measured feature, μ is the location parameter and σ is the scale parameter of the data in the logarithmic scale. For that reason we used expected value ev of variable x as a characteristic value for each phoneme class. The expected value is defined as:

TABLE II

The relation of the applied notation of phonemes (CORPORA) with SAMPA and IPA notation. The frequency of appearance of phonemes' realizations in databases Db1 (isolated sentences) and Db2 (continuous speech).

Notation			Frequency of appearance [%]	
CORPORA	SAMPA	IPA	I	II
a	a	a	8.5	8.82
o	O	ɔ	7.61	8.61
e	e	ɛ	6.93	9.99
i	i	i	4.4	4.33
m	m	m	4.26	2.88
n	n	n	4.22	4.43
y	I	ɨ	4.08	4.61
u	u	u	3.58	3.72
j	j	j	3.57	4.34
r	r	r	3.21	3.39
k	k	k	3.14	3.56
l_	W	w	3.03	1.11
t	t	t	2.86	4.57
s	s	s	2.77	2.77
l	l	l	2.75	2.19
w	v	v	2.71	4.28
p	p	p	2.52	3.46
a_	o w∼	ɔ̃w̃	2.29	0.99
ni	n′	ɲ	2.1	2.73
d	d	d	2.03	2.56
b	b	b	2.02	1.42
h	x	x	1.92	0.87
rz	Z	ʒ	1.89	1.27
si	s′	ɕ	1.88	1.34
z	z	z	1.86	2.56
c	ts	t͡s	1.69	1.32
sz	S	ʃ	1.54	1.43
cz	tS	t͡ʃ	1.5	1.07
ci	ts′	t͡sʲ	1.46	1.07
g	g	g	1.45	1.45
f	f	f	1.36	1.3
zi	z′	ʑ	1.24	0.11
dzi	dZ	d͡ʑ	1.12	0.87
e_	e j∼	ɛ̃ŋ	0.92	0
drz	dz′	d͡zʲ	0.8	0.07
dz	dz	d͡z	0.52	0.29
N	N	ŋ	0.28	0.23

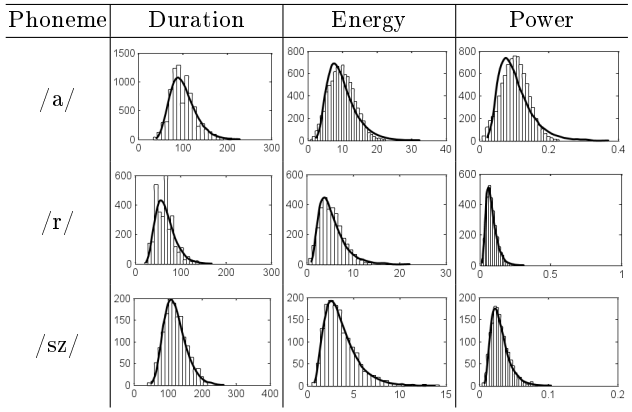
$$ev(x) = \mathrm{e}^{\mu(x) + \frac{\sigma^2(x)}{2}}. \tag{5.5}$$

The results are presented in Table and the visualization of Polish phonemes on a duration-energy two-dimensional space is shown in Fig. 1.

Using the expected values of each phoneme class, the normalization is performed by dividing each phoneme value by the expected value for the phoneme class, according to the equations

TABLE III

The distribution of duration, energy and power parameters for 3 example phonemes with lognormal probability density function added.



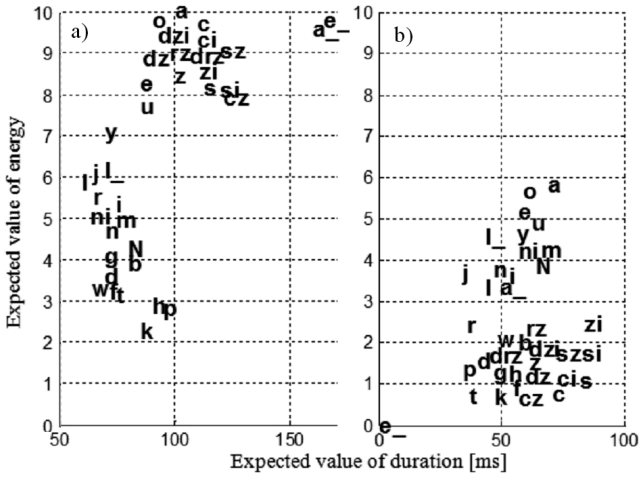


Fig. 1. The map of Polish phonemes on a duration-energy two-dimensional space for a) Db1, b) Db2.

$$\widetilde{D}_i = \frac{D_{m,i}}{ev\,(D_m)}, \tag{5.6}$$

$$\widetilde{E}_i = \frac{E_{m,i}}{ev(E_m)}, \tag{5.7}$$

$$\widetilde{P}_i = \frac{P_{m,i}}{ev(P_m)}. \tag{5.8}$$

There are several algorithms for the pitch contour extraction. The most popular are: the cepstral method, the autocorrelation, the zero-crossing, and the subharmonic-to-harmonic ratio. The detailed comparative evaluation can be found in [44]. For F0 tracking we used the YAAPT algorithm [45] because of its robustness. The algorithm is based on normalized cross-correlation and dynamic programming methods. We also benefited from Reichel and Mády works [46] on the issue of F0 contour parameterization.

TABLE IV

The expected values of duration, energy and power for phoneme classes.

CORPORA notation	Expected value of					
	duration [ms]		energy		power	
	Db1	Db2	Db1	Db2	Db1	Db2
a	100	69	8.936	5.812	0.092	0.084
o	90	59	8.269	5.662	0.098	0.095
e	85	57	2.857	5.145	0.037	0.09
i	75	51	5.313	3.636	0.072	0.076
m	75	63	3.359	4.198	0.046	0.07
n	70	50	3.136	3.634	0.044	0.076
y	70	58	3.396	4.563	0.049	0.077
u	85	56	2.257	4.645	0.029	0.082
j	65	34	5.002	3.67	0.076	0.107
r	65	36	6.088	2.397	0.098	0.068
k	85	47	7.709	0.649	0.095	0.016
l_	70	44	6.135	4.545	0.09	0.106
t	75	37	4.938	0.691	0.064	0.02
s	115	82	9.018	1.063	0.092	0.014
l	60	44	5.828	3.323	0.092	0.083
w	70	50	4.026	1.758	0.062	0.036
p	95	36	9.211	1.485	0.1	0.045
a_	160	49	9.569	3.535	0.089	0.073
ni	65	57	5.495	4.224	0.085	0.076
d	70	42	3.611	1.554	0.055	0.046
b	80	57	4.095	1.983	0.052	0.041
h	90	55	2.808	1.168	0.034	0.024
rz	100	60	10.009	2.323	0.104	0.04
si	120	83	8.102	1.707	0.082	0.023
z	100	60	8.567	1.585	0.094	0.028
c	110	71	9.698	0.744	0.101	0.012
sz	120	72.5	8.089	1.703	0.082	0.025
cz	120	51	8.118	1.128	0.085	0.022
ci	110	73	8.955	1.109	0.092	0.015
g	70	48.5	4.837	1.341	0.068	0.034
f	70	55	7.089	0.863	0.1	0.017
zi	110	84	9.165	2.423	0.095	0.03
dzi	95	61	9.781	1.83	0.108	0.033
e_	165	0	9.762	0	0.09	0
drz	105	46	8.936	1.279	0.092	0.023
dz	100	57	8.922	1.079	0.092	0.02
N	80	64	3.905	4.093	0.05	0.068

6. Statistical analysis and modeling of prosodic features correlated with sentence boundaries

Having calculated the ratio of each phone duration in relation to its expected value for the phoneme class, we investigated how the phonemes are lengthened at the end of a phrase.

The white histogram (see Fig. 2) shows distribution of relative durations for all phonemes, while the black one — distribution of relative durations of the last phonemes

in the sentences. The majority (88.5% for Db1 and 91.7% for Db2) of the last phonemes in the phrase were much longer than average. The mean values of their duration ratio were 1.54 (Db1) and 1.85 (Db2), which indicates that an average realization of the last phoneme in a phrase was at least about 50% longer than an average realization of a phoneme in the whole database.

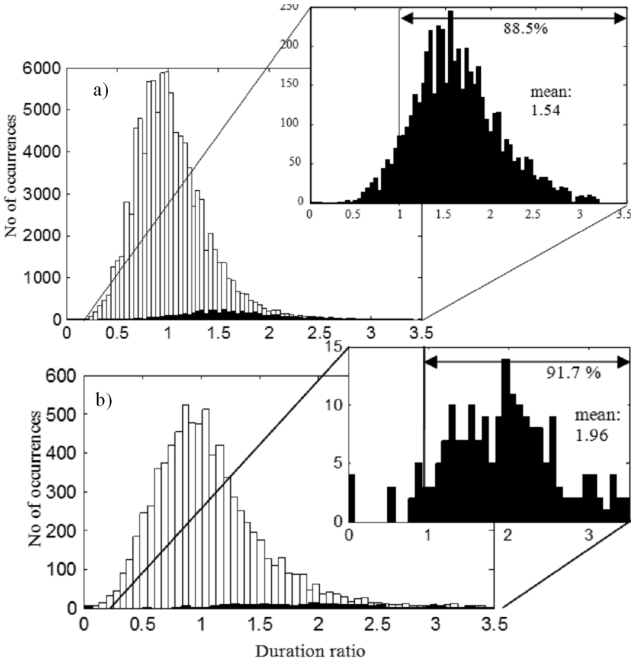


Fig. 2. The distribution of the last phonemes' duration ratios (black) compared to all phonemes' duration ratios (white): a) Db1, b) Db2.

For the Database 2, it was possible to look for the tendencies in places of full stops and commas separately. Figure 3 shows that distribution of the last phonemes reveals the same lengthening feature

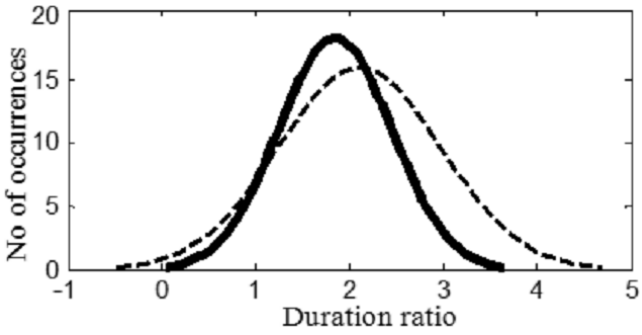


Fig. 3. The distribution of duration ratios of the last phonemes before a comma (dotted line) and before a full stop (solid line).

On the basis of the distribution characteristics, a probability model can be constructed. It defines conditional probability that the phoneme is the last one in a sentence, given its relative duration (Fig. 4). The longer the duration ratio of the phoneme, the higher the probability that the phoneme is located at the end of a phrase. We obtained the best approximation (root mean square error: 0.03 for Database 1 and 0.4 for Database 2) of this relationship using the Gaussian model

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{6.1}$$

The relation between the duration ratio and the probability that the phoneme is the last one in the phrase are illustrated in Fig. 4 and described by the equations

$$\text{Database I: } P(\widetilde{D}_i)=0.69\exp\left(-\frac{(\widetilde{D}_i-2.55)^2}{0.81}\right), \tag{6.2}$$

$$\text{Database II: } P(\widetilde{D}_i)=0.29\exp\left(-\frac{(\widetilde{D}_i-2.60)^2}{0.72}\right). \tag{6.3}$$

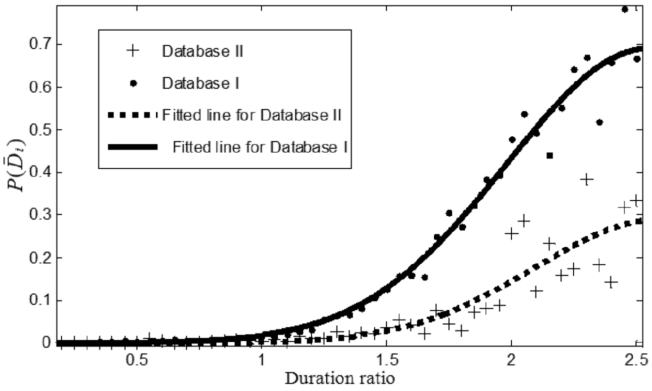


Fig. 4. The probability that a phoneme with a given duration ratio is the last one in a phrase.

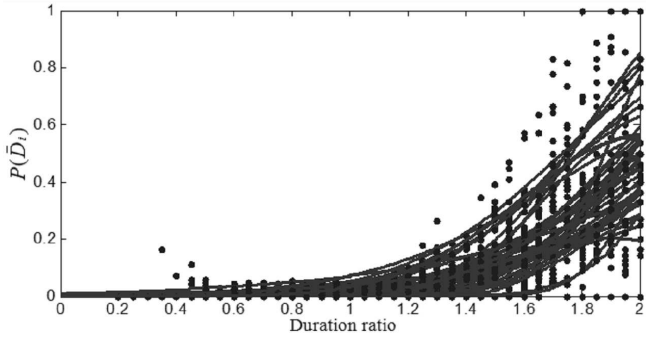


Fig. 5. The comparison of probability that a phoneme with a given duration ratio is the last one in a phrase depending on different speakers.

Nevertheless, the tendency of lengthening the last phonemes in a phrase occurs in varying degrees depending on different speakers. As it was shown in Fig. 5, for some of them it is clearly marked while for others just slightly signalized. We describe it in detail in [47].

The next way to describe the phenomenon of the preboundary lengthening is constructing plots of relative duration ratios within a sentence (examples are presented

in Fig. 6). We approximate the tendency of increasing phoneme length with a linear regression (the dotted line on the plots) and take its first coefficient a (tangent of its slope angle) as a determinant if the phonemes were lengthened towards the end of the phrase and how strong this tendency was. To make the result phrase length independent, we used the linear regression coefficient of the last 10 phonemes for further calculations of each phrase. The distribution of the coefficients obtained from phrases in both databases is presented in Fig. 7. For the majority of them (89% for Db1 and 93% for Db2), the coefficient a was greater than zero, meaning that the preboundary lengthening occurred. The average values of the coefficient were 0.05 (Db1) and 0.08 (Db2). If we take into account last 5 phonemes instead of 10, we achieve an average 0.166 (Db1), 0.174 (Db2) and a positive value for 93% of sentences for Db1 and 81% for Db2.

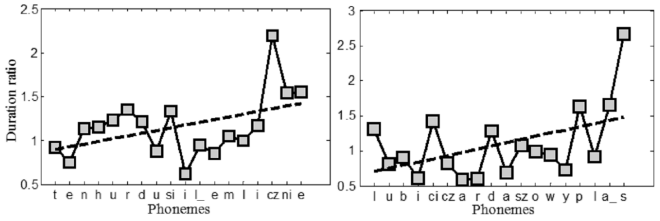


Fig. 6. Example phoneme durations within a sentence (grey squares) with the dynamic average of the duration ratios (dotted line).

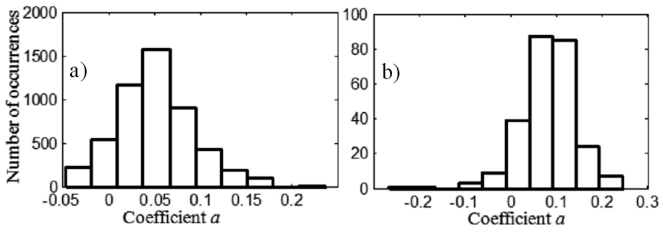


Fig. 7. Distribution of a slope coefficient for linear regression measured for last 10 phonemes of each sentence: (a) Db1, (b) Db2.

In another article [36], we described how the observation of a dynamic average of these values and its standard deviation can be applied to the automatic sentence boundary detection.

To investigate how energy and power change at the end of phrases, we undertook the analysis steps similar to those described for the duration parameter. As it was expected and widely known [14–16, 26], both energy and power decrease towards the end of the sentence. Therefore, the majority of the last phonemes are localized on the left of all phonemes' histograms (Fig. 8 and Fig. 9). At the same time, probability density models show decreasing probability of the end of the sentence along with increasing ratios of energy or power (Fig. 10) according to the equations

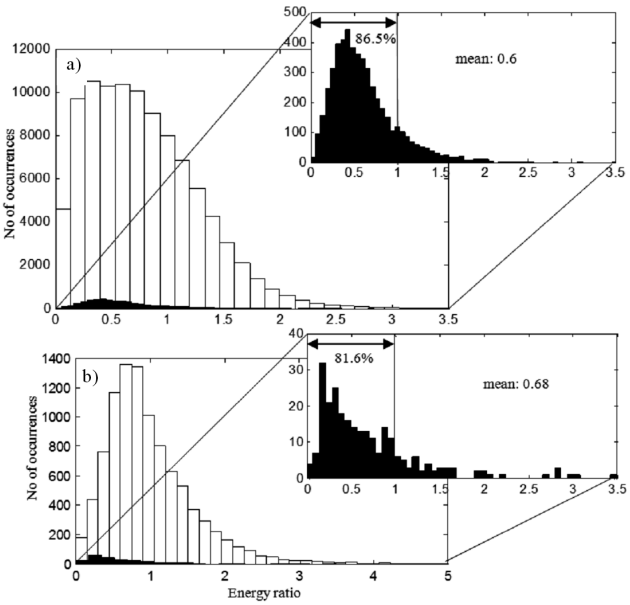


Fig. 8. The distribution of all phonemes' energy ratios (white) and last phonemes' energy ratios (black): (a) Db1, (b) Db2.

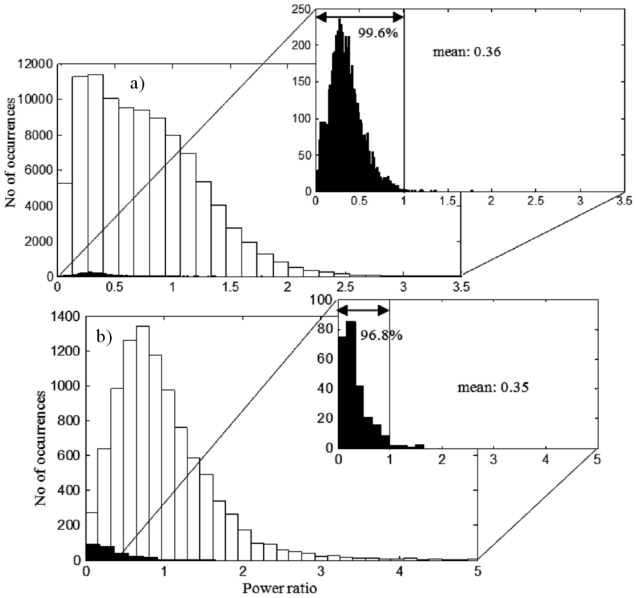


Fig. 9. Distribution of all phonemes' power ratios (white) and last phonemes' power ratios (black): (a) Db1, (b) Db2.

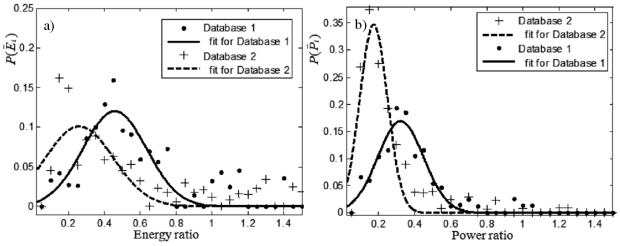


Fig. 10. Probability that a phoneme with a given energy (left) or power (right) ratio is the last one in the sentence: (a) Db1, (b) Db2.

$$\text{Database 1: } P(\widetilde{E}_i)=0.12\exp\left(-\left(\frac{\widetilde{E}_i+0.46}{0.25}\right)^2\right), \quad (6.4)$$

$$\text{Database 2: } P(\widetilde{E}_i)=0.10\exp\left(-\left(\frac{\widetilde{E}_i+0.26}{0.27}\right)^2\right), \quad (6.5)$$

$$\text{Database 1: } P(\widetilde{P}_i)=0.17\exp\left(-\left(\frac{\widetilde{P}_i-0.33}{0.18}\right)^2\right), \quad (6.6)$$

$$\text{Database 2: } P(\widetilde{P}_i)=0.35\exp\left(-\left(\frac{\widetilde{P}_i-0.17}{0.11}\right)^2\right). \quad (6.7)$$

Similarly to the previous computation for the duration parameter, linear regression coefficients were gathered in Table V.

TABLE V

Energy and power parameters $\frac{\text{mean value}}{\text{negative rate}}$ (linear regression for last k phonemes).

	Db1		Db2	
	last 10	last 5	last 10	last 5
energy	−0.0065	−0.0423	−0.0186	−0.0436
	53%	73%	64%	70%
power	−0.0274	−0.1013	−0.0559	−0.0959
	80%	87%	87%	89%

For better visualization of the relation between duration and energy of phonemes located in the endings of sentences, maps on two-dimensional space were prepared (Fig. 11). The last phonemes create a cluster in the bottom of the plot (high duration, low energy).

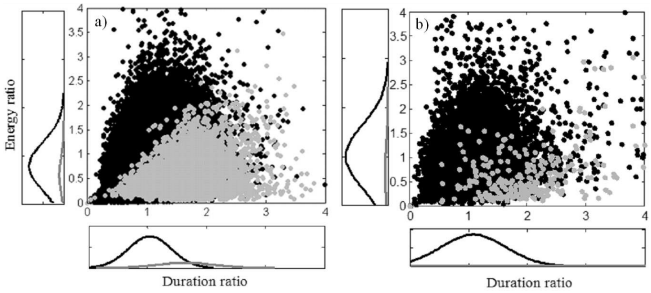


Fig. 11. The map of phonemes (black) and end-of-sentence phonemes (grey) in the duration–energy space: (a) Db1, (b) Db2.

Ends of statements should be characterized by the pitch contour cadence. We have investigated which parameters of pitch describe the phenomenon quantitatively.

As a result of extracting pitch contours with the YAAPT algorithm, we have obtained voiced and unvoiced regions of speech signal. For every voiced region, parameters of mean, minimum, maximum, amplitude, standard deviation and tangent of linear regression were calculated. The same parameters were computed globally, for the entire utterance. The visualization of contour alignment is presented in Fig. 12.

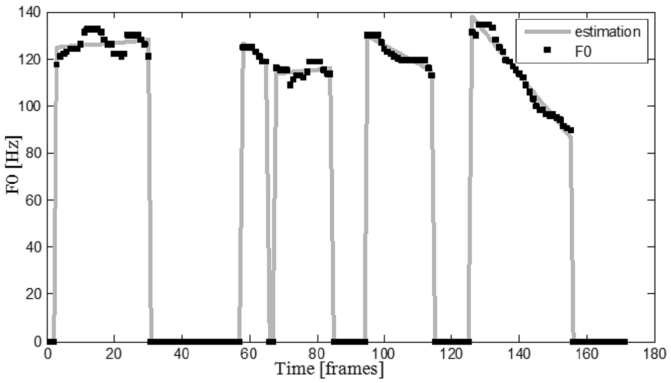


Fig. 12. An example of a pitch contour within a sentence, frames: 10 ms.

Table VI (at the end) contains the most significant numerical values found to be good descriptors of the tendency of pitch decrease in the ending of an informative sentence.

All the differences found for acoustic features at the sentence endings were statistically significant (at the significance level 0.05).

7. Statistical analysis and modeling of prosodic features correlated with accents

According to the works reviewed in Sect. 3, accented vowels in Polish should be lengthened in comparison to non-accented ones. In order to determine the scale of the tendency, we investigated average duration ratios of accented vowels. Although the differences are statistically significant ($p < 0.01$ at the significance level 0.05), the distribution of accented vowels (Fig. 13) is comparable with distribution of all vowels, and the average value is only slightly greater than average for all vowels.

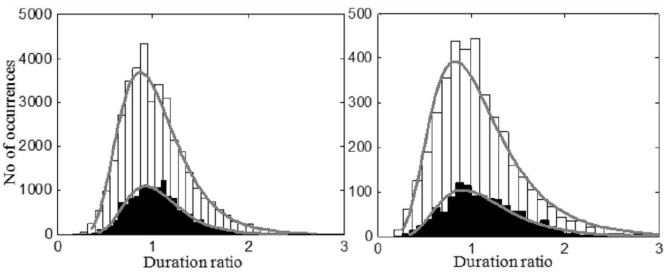


Fig. 13. The distribution of accented vowels' duration ratios (black histogram) against distribution of all vowels (white histogram): (a) Db1, (b) Db2.

The case study of vowel /a/ shows that although the mean for accented /a/ is 107.5 ms and for non-accented

is 93 ms, the difference is too small to differentiate them, as their histograms overlap (Fig. 14).

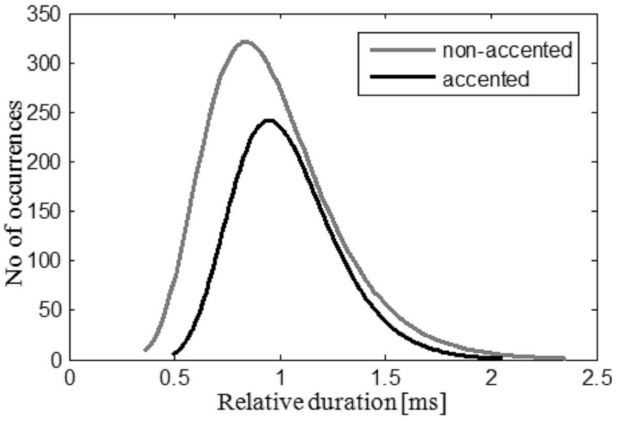


Fig. 14. The models of distribution of accented and non-accented /a/ phoneme.

Since the duration parameter does not seem to be an appropriate measure of accent, the energy and power parameters are much more characteristic. In accordance with phonological description, energy and power are good (statistically significant) descriptors of vocal stress on accented syllable. As we can observe in Fig. 15, accented vowels' histogram is moved into the right of all vowels. The probability that the vowel with a given energy ratio is stressed grows with an increase of the energy ratio, according to the Gaussian model

$$\text{Database 1: } P(\widetilde{E}_i) = 0.71 \exp\left(-\left(\frac{\widetilde{E}_i + 2.12}{1.35}\right)^2\right), \quad (7.1)$$

$$\text{Database 2: } P(\widetilde{E}_i) = 0.38 \exp\left(-\left(\frac{\widetilde{E}_i + 1.64}{1.56}\right)^2\right), \quad (7.2)$$

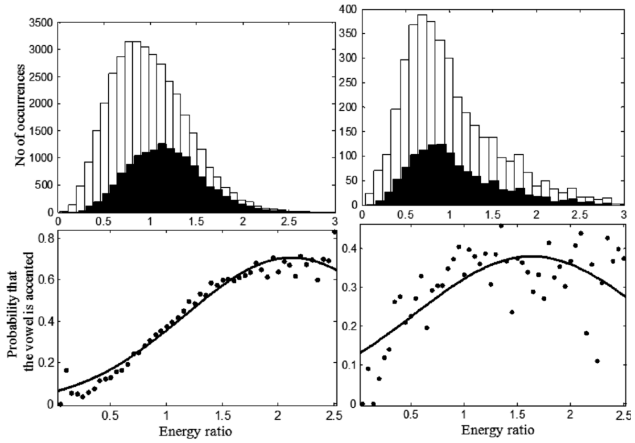


Fig. 15. The distribution of energy ratios for accented and non-accented vowels (left) and the probability model describing probability that the vowel is accented, given its energy ratio (right).

For accents we should observe a local increase of a pitch. We measured the regularity of the tendency by comparing a maximum value of accented vowels with a baseline pitch level in a sentence (the baseline was calculated as a mean value of all non-zero pitch values). We found that in the Database 1 the average difference between the local F0 in an accented vowel and the baseline was 18.4 Hz, while in the Database 2 it was 6.5 Hz.

We also compared mean F0 values and F0 maximum for accented and non-accented vowels from the same class. The results (see Table VII) confirm the role of pitch in accent signalization (at the significance level 0.05). For every vowel we noted at least a slight increase of F0 in the place of the main word accent.

TABLE VII

The parameters describing local increase of F0 on accented vowels.

Vowel	Non-accented vowel mean F0 [Hz]		Accented vowel mean F0 [Hz]		Increase on accent [Hz]	
	Db1	Db2	Db1	Db2	Db1	Db2
'a'	159	172	170	174	+11	+4
'a_'	154	173	168	210	+14	+37
'e'	159	173	172	174	+13	+1
'e_'	153	–	175	–	+22	–
'o'	156	165	165	169	+9	+4
'u'	163	169	178	183	+15	+14
'i'	152	168	164	169	+12	+1
'y'	168	167	183	179	+15	+12
Vowel	Non-accented vowel max F0 [Hz]		Accented vowel max F0 [Hz]		Increase on accent [Hz]	
	Db1	Db2	Db1	Db2	Db1	Db2
'a'	173	178	183	180	+10	+2
'a_'	172	178	184	213	+12	+35
'e'	176	179	186	180	+10	+1
'e_'	167	–	191	–	+24	–
'o'	177	172	185	175	+8	+3
'u'	180	174	196	188	+16	+14
'i'	175	173	185	175	+10	+2
'y'	181	173	196	184	+15	+11

8. Discussion and conclusion

The corpora that we used differ in several aspects. The content of sentences in the Database 1 was designed in order to obtain the best possible distribution of Polish diphones. Therefore, semantically they often make no sense; they are unnatural and contain obsolete vocabulary or rare sequences of phones, while continuous texts from Database 2 are natural and more similar to daily speech. In [48] a comparison of phonemes' frequency of appearance rankings (based on the results obtained by Jassem, Łobacz, Rocławski, and Steffen) was made and it was concluded that Polish phonemes' inventory structure

is rather constant and independent from spoken or written language. Only for Database 2 we obtained a very similar distribution and proportion of phonemes (with /e/, /a/, /o/, /j/, /t/, /y/, /m/ as the most frequent phonemes, see Table II).

The artificial style of Database 1 determines the specific way of pronunciation (slow, careful and precise). It is reflected in average values of phones' duration as well as energy (Table IV), which are significantly higher than the results from Database 2 (it is also illustrated in Fig. 1). The differences of relative duration, energy and power within a phrase are smaller in the Database 1. Regression models indicate that duration lengthening towards the ending of a sentence (Fig. 2 and Fig. 7) is weaker than in Db2. Additionally, energy and power decrease were observed (Table V). The opposite tendency was found for a local rise of F0 in places of accented vowels — it is clearer for the Database 1 (Fig. 15 and Table VII).

Although the 2-dimensional phonemes' maps (see Fig. 1) vary in duration and energy values, their proportions remain similar and some general tendencies can be found. Polish phonemes make groups corresponding to their phonological nature. Vowels are placed on the top (the highest energy). Among consonants, sonorants have low duration and high energy. Nasals form a group of high duration and low energy, while plosives and fricatives have both low duration and energy.

Concerning the changes of duration, energy, power and pitch in the ending of phrases we confirmed that all these parameters are statistically meaningful, which is consistent with the previous results obtained by Demenko [26] and Klessa [4, 8]. Prosodic cues are natural discourse demarcation indicators complementary to speech content. They convey structural, semantic, and functional information and most of them are resistant to communication channel characteristics changes. In [49] it was shown that over 65% of syntactic boundaries were coded in prosodic information. Saloni and Świdziński [50] made a review of the criteria of sentence extraction. Among others, phonological criterion is considered: sentences are formed between long pauses and they are units with closed intonational contour with rising or falling cadence. Our results for prosodic cues characteristic for sentence boundaries are also similar to those obtained for other languages [14, 16]. Further steps will include the development of algorithms for automatic punctuation annotation in speech transcripts.

The nature of accent in Polish remains disputable. In our research we have found only slight lengthening of accent vowels in both databases (similarly to Jassem [29] and Malisz [24], whereas Klessa [4] found a slight increase of duration depending on the region the speaker is from). Increase of energy as well as a local rise of pitch were meaningful and statistically significant. It would suggest that Polish accent has properties of both the dynamic accent and the pitch accent, which remains in consistency with [26, 30] and recent results by Malisz [24]. It is also partially concordant with Jassem results [29].

However, in our work only basic prosodic parameters were analysed. Another weakness of our study is that we did not differentiate particular types of accents. This issue requires more advanced research using a wider range of parameters (like [51] for Slovak and Hungarian or [52] for French) and more advanced modelling tools (e.g. [53] or [54] for English and others), which is also planned to be performed in future works.

The results of the investigation are to be applied in designing algorithms for the automatic punctuation detection and will be used for language modeling in automatic speech recognition systems.

TABLE VI

The parameters describing a pitch cadence in endings of informative sentences.

Parameter	Description/interpretation	Database 1	Database 2
tgF0_of_the_last_voiced_fragment	how fast pitch is decreasing in the ending of the sentence	mean: –1.14 negative for 91%	mean: –0.72 negative for 62%
tgF0_of_the_full_sentence	how fast pitch is decreasing within the sentence	mean: –0.36 negative for 97%	mean: –0.2 negative for 81%
min_tgF0_of_all_voiced_regions	how often pitch decrease was the greatest in the last voiced region	38%	38%
F0_amplitude_of_the_last_voiced_region	how much pitch decreases within the last voiced region	48 Hz	14 Hz
F0_amplitude_for_the_last_voiced_region	how much pitch decreases within the sentence	98.4 Hz	23 Hz
is_last_meanF0_the_smallest	how often the mean pitch of the last voiced region is smaller than the mean global pitch	96%	94%
is_last_minF0_the_smallest	how often the minimal pitch of the last voiced region is the smallest in the sentence	77%	46%
min_F0_in_the_last_voiced_region_in_reference_to_global_mean [Hz]	how much is the minimum pitch of the last voiced segment smaller than the mean pitch for the sentence	46 Hz	37 Hz
min_F0_in_the_last_voiced_region_in_reference_to_global_mean [%]	how much is the minimum pitch of the last voiced segment smaller than the mean pitch for the sentence	–27%	–22%

Acknowledgments

The research was funded by the National Science Centre allocated on the basis of the decision DEC-2011/03/D/ST6/00914.

References

[1] G. Demenko, M. Szymański, R. Cecko, E. Kuśmierek, M. Lange, K. Wegner, K. Klessa, M. Owsianny, *Acta Phys. Pol. A* **121**, A86 (2012).

[2] M. Ziółko, J. Gałka, B. Ziółko, T. Jadczyk, D. Skurzok, M. Mąsior, in: *Proc. Interspeech, Florence (Italy)*, Eds.: P. Cosi, R. De Mori, G. Di Fabbrizio, R. Pieraccini, ISCA, Florence 2011, p. 3315.

[3] B. Ziółko, M. Ziółko, in: *Lect. Notes Artif. Intell. 6562*, Ed. Z. Vetulani, Springer-Verlag, Berlin 2011, p. 105.

[4] K. Klessa, in: *Speech and Language Technology, Special Issue dedicated to Wiktor Jassem*, Eds. D. Gibbon, D. Hirst, N. Campbell, PTFon, Poznan 2011/2012, p. 94.

[5] W. Wszołek, A. Izworski, G. Izworski, *Acta Phys. Pol. A* **123**, 995 (2013).

[6] J. Imiołczyk, I. Nowak, G. Demenko, *Arch. Acoust.* **19**, 2 (1994).

[7] K. Klessa, D. Śledziński, in: *Speech and Language Technology*, Ed. G. Demenko, Polish Phonetic Association, Poznań 2008, p. 87.

[8] K. Francuzik (Klessa), M. Karpiński, J. Kleśta, E. Szalkowska, *Speech Prosody Proc., Aix-en-Province (France)*, Eds.: B. Bel, I. Marlien, ProSig and Universite de Provence Laboratoire Parole et Langage, Aix-en-Provence 2002.

[9] SJP, *Polish Spelling and Punctuation Rules*, PWN, Warsyawa 2010 (in Polish).

[10] E. Polański, M. Szopa, *Handbook of Polish Punctuation*, MAC Edukacja S.A., Kielce 2004 (in Polish).

[11] T. Karpowicz, *Culture of Polish: Pronunciation, Orthography, Punctuation*, PWN, Warszawa 2012 (in Polish).

[12] F. Przyłubski, *Poradnik Językowy* no. 8, 1953 (in Polish), p. 11.

[13] E. Łuczynski, *Współczesna interpunkcja polska. Norma a uzus.*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk 1999 (in Polish).

[14] E. Shriberg, A. Stolcke, D. Hakkani-Tur, G. Tur, in: *Speech Communication*, Vol. 32, Eds.: S. Renals, T. Robinson, Elsevier, 2000, p. 127.

[15] J. Kolar, J. Svec, J. Psutka, in: *SPECOM'2004*, SPIIRAS, Saint-Petersburg 2004, p. 319.

[16] D. Baron, E. Shriberg, A. Stolcke, in: *Proc. Int. Conf. on Spoken Language Processing*, Denver 2002, p. 949.

[17] Z. Pawłowski, *Voice Emission — Structure, Function, Diagnostics, Pedagogy*, Wydawnictwo Salezjańskie, Warszawa 2008 (in Polish).

[18] A. Wagner, Ph.D. Thesis, Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza, Poznań 2008.

[19] A. Wagner, in: *Proc. Speech Prosody, Chicago (USA)*, 2010.

[20] G. Demenko, A. Wagner, *Arch. Acoust.* **32**, 25 (2007).

[21] B.A. Hockey, Z. Fagyal, in: *Proc. 14th Int. Congress of Phonetics Sciences, San Francisco*, Eds. K. Laitakari, M. Luotonen, San Francisco 1999, p. 313.

[22] M. Shepherd, in: *USC Working Papers in Linguistics 4*, Ed. M. Shepherd, University of Southern California, 2011, p. 1.

[23] E. Grabe, M. Karpiński, in: *Proc. 15th Int. Congress of Phonetic Sciences, Barcelona (Spain)*, Eds.: M. J. Solé, D. Recasens, J. Romero, ICPhS, Barcelona 2003, Vol. 1, p. 1061.

[24] Z. Malisz, P. Wagner, in: *Rhythm, Melody and Harmony in Speech. Studies in Honour of Wiktor Jassem*, Eds. D. Gibbon, D. Hirst, N. Campbell, Speech and Language Technology, Poznań 2012, p. 105.

[25] M. Wypych, Ph.D. Thesis, PAN, Warszawa 2011.

[26] G. Demenko, *Analysis of Polish Suprasegmentals for Speech Technology*, Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza, Poznań 1999 (in Polish).

[27] L. Frąckowiak-Richter, in: *Speech Analysis and Synthesis III*, Ed. W. Jassem, PWN, Warszawa 1973, p. 87.

[28] D. Ostaszewska, J. Tambor, *Phonetics and Phonology of Modern Polish Language*, PWN, 2000 (in Polish).

[29] W. Jassem, *Accent of Polish*, Ossolineum, Wrocław 1962 (in Polish).

[30] B. Wierzchowska, *Polish Pronunciation*, PWN, Warszawa 1971 (in Polish).

[31] B. Rocławski, *An outline of phonology, phonotactics and phonostatistics of contemporary Polish*, Gdańsk University, Gdańsk 1976 (in Polish).

[32] L. Richter, *Speech Analysis and Synthesis (Warszawa)* **3**, 87 (1973).

[33] M. Dłuska, *Polish Prosody*, Państwowe Wydawnictwo Naukowe, Warszawa 1976 (in Polish).

[34] G. Dogil, in: *Word Prosodic Systems in the Languages of Europe*, Ed. H. van der Hulst, Mouton de Gruyter, Berlin 1999.

[35] H. Mackiewicz-Krassowska, *Intonation of English and Polish Declarative Sentences*, PSiCL2, 1974.

[36] M. Steffen-Batogowa, *Accentual Structure of Polish*, PWN, Warsaw 2000 (in Polish).

[37] D. Oliver, R. Clark, in: *Proc. Interspeech'2005 — Eurospeech, 9th Europ. Conf. on Speech Communication and Technology, Lisbon (Portugal)*, Ed. I. Trancoso, ISCA, Lisbon 2005, p. 1965.

[38] L. Newlin-Łukowicz, *Phonology***29**, 271 (2012).

[39] S. Grocholewski, in: *Fifth European Conference on Speech Communication and Technology, EUROSPEECH*, Eds.: G. Kokkinakis, N. Fakotakis, E. Dermatas, ISCA, Rhodes 1997, p. 1735.

[40] M. Igras, B. Ziółko, M. Ziółko, in: *Int. Conf. on Simulation and Modeling Methodologies, Technologies and Applications, Reykjavik*, ICETE, Reykjavik 2013.

[41] B. Klebanowska, *Phonological Interpretation of Phonetic Phenomena in Polish, with Exercises*, Warsaw University, Warsaw 2007 (in Polish).

[42] M. Bartkowiak, A. Borowicz, J. Bułat, P. Chołda, M. Domański, K. Duda, P. Dymarski, M. Grega, L. Janowski, P. Korohoda, D. Kól, M. Leszczuk, W. Ludwin, R. Makowski, A. Pach, Z. Papir, M. Parfieniuk, W. Półchłopek, J. Rachwalski, R. Rumian, P. Świętojański, P. Turcza, R. Wielgat, J. Wszołek, T. Zieliński, *Digital Signal Processing in Telecommunications: Fundamentals — Multimedia — Transmission*, Rds.: T. Zieliński, P. Korohoda, R. Rumian, PWN, Warszawa 2014 (in Polish).

[43] B. Ziółko, M. Ziółko, *Speech Processing*, Wydawnictwa AGH, Kraków 2011 (in Polish).

[44] A. de Cheveigné, H. Kawahara, in: *INTERSPEECH*, Eds.: P. Dalsgaard, B. Lindberg, H. Benner, ISCA, Aalborg 2001, p. 2451.

[45] S.A. Zahorian, H. Hongbing, *J. Acoust. Soc. Am.* **123**, 4559 (2008).

[46] U.D. Reichel, K. Mády, in: *Elektronische Sprachverarbeitung. Studientexte zur Sprachkommunikation*, Vol. 65, Ed. P. Wagner, TUDpress, Dresden 2013, p. 223.

[47] M. Igras, B. Ziółko, M. Ziółko, in: *Pacific Voice Conference (PVC)*, Eds. B. Ziółko, J. Grzybowska, IEEE, Kraków 2014, p. 1.

[48] J. Sambor, in: *Contemporary Polish*, Wyd. UMCS, Lublin 2014, p. 503 (in Polish).

[49] M.L. Fach, in: *Proc. Europ. Conf. on Speech Communication and Technology*, ISCA, Budapest 1999, p. 527.

[50] Z. Saloni, M. Świdziński, *Syntax of Contemporary Polish*, PWN, Warszawa 1998 (in Polish).

[51] K. Mády, U.D. Reichel, Š. Beňuš, in: *Proc. Speech Prosody*, Eds.: N. Campbell, D. Gibbon, D. Hirst, Dublin 2014, p. 752.

[52] A. Lazaridis, J.P. Goldman, M. Avanzi, P.N. Garner, in: *Nouveaux cahiers de linguistique francaise*, 2014.

[53] F. Biadsy, Ph.D. Thesis, Columbia University, 2011.

[54] S. Gaikwad, B. Gawali, K.V. Kale, *Int. J. Comput. Appl.* **63**, 3 (2013).