PAN
ARCHIVES OF ACOUSTICS
Vol. **43**, No. 2, pp. 253–262 (2018)

DOI: 10.24425/122373

Development of the Polish Speech Test Signal and its Comparison with the International Speech Test Signal

Dorota HABASIŃSKA, Ewa SKRODZKA, Edyta BOGUSZ-WITCZAK

Institute of Acoustics
Faculty of Physics
Adam Mickiewicz University in Poznań
Umultowska 85, 61-614 Poland; e-mail: afa@amu.edu.pl

(*received October 12, 2017; accepted January 5, 2018*)

The aim of this study was to create a single-language counterpart of the International Speech Test Signal (ISTS) and to compare both with respect to their acoustical characteristics. The development procedure of the Polish Speech Test Signal (PSTS) was analogous to the one of ISTS. The main difference was that instead of multi-lingual recordings, speech recordings of five Polish speakers were used. The recordings were cut into 100–600 ms long segments and composed into one-minute long signal, obeying a set of composition rules, imposed mainly to preserve a natural, speech-like features of the signal. Analyses revealed some differences between ISTS and PSTS. The latter has about twice as high volume of voiceless fragments of speech. PSTS's sound pressure levels in 1/3-octave bands resemble the shape of the Polish long-term average female speech spectrum, having distinctive maxima at 3–4 and 8–10 kHz which ISTS lacks. As PSTS is representative of Polish language and contains inputs from multiple speakers, it can potentially find an application as a standardized signal used during the procedure of fitting hearing aids for patients that use Polish as their main language.

Keywords: Polish Speech Test Signal (PSTS); International Speech Test Signal (ISTS); hearing aids fitting; language.

1. Introduction

At the beginning of the 21st century, digital hearing aids became prevalent over analogue ones on the hearing instruments world market. The digital ones are equipped with various advanced digital systems designed to improve their performance, especially by the means of optimal processing of speech signal. Nevertheless, until 2012 there were no standard procedures to test the actual hearing aids performance in the presence of speech signal. With the aim of filling this gap, a new standard method for charactersising signal processing in hearing aids with a speech-like test signal, ISTS (International Speech Test Signal), has been introduced (IEC 60118-15, 2012; Holube *et al.*, 2010). ISTS was developed using female speech recordings in six languages: Mandarin, Arabic, English, Spanish, French and German. This particular choice of languages including languages from Sino-Tibetan, Semitic, Germanic and Romance groups was not justified explicitly, but at least in the case of Mandarin, Arabic, Spanish and English a percentage of world's population using these languages must have played a role in that. However, ISTS does not contain characteristic features of languages from other groups, e.g. Indo-Aryan, Japonic, Slavic or Turkic.

In ISTS fragments of recordings of six languages were composed in a rigorous and diligent manner in order to preserve all the key features of human speech. The signal itself allows for repeatable measurements and therefore it has been proven successful as an international standard and as an input signal applied during hearing aids' tests. Since its release ISTS has found also other applications, e.g. in the process of fitting hearing aids for individual patients (Dworsack-Dodge, Switalski, 2012); Kossek, Dworsack-Dodge, 2010). In this context, it is worth noting that the limited linguistic content of ISTS implies that the signal may not be representative of any given language (Holube *et al.*, 2010).

Due to the fact it is assumed here, that for patients using specific languages, fitting a hearing with ISTS might not provide optimal results. One of possible reasons for that might be a specific phonemic content of the patient's language (Chasin, Hockley, 2013). The method that uses ISTS in hearing aid fitting process is based on percentile analysis. Details of the analysis are described in the IEC standard (IEC 60118-15, 2012) while guidelines for fitting practices can be found in (Dworsack-Dodge, Switalski, 2012; Kossek, Dworsack-Dodge, 2010).

Importantly, ISTS development procedure was based on the assumption that long-term average speech spectra of particular languages do not differ significantly (Byrne *et al.*, 1994) and for that reason ISTS's long-term average spectrum mirrors exactly the shape of the International Long-term Averaged Speech Spectrum (ILTASS). However, each vowel or consonant manifests some uniqueness in the spectrum, and since different languages use different sets and incidence of speech sounds, some researchers have reckoned that the shape of LTASS might be language-specific (Noh, Lee, 2012; de Boysson-Bardies *et al.*, 1986).

As for the Polish speech – apart from the popular opinion on its specific sound (often commonly described as rustling or whistling) – there are scientific reports on its spectral uniqueness:

- Polish consonant phonemes include the series of affricates: *c* /t͡s/, *dz* /d͡z/ (alveolar), *cz* /t͡ʂ/, *dż* /d͡ʐ/ (retroreflex), *ć* /t͡ɕ/, *dź* /d͡ʑ/ (palatal), palatal consonants: *ń* /ɲ/, *ś* /ɕ/, *ź* /ʑ/, *j* /j/, two palatalized plosives /ki/, /gi/ and one palatalized fricative /xi/ (Jassem, 2003). All of them exhibit significant amount of energy in high frequency spectrum region.
- Ozimek *et al.* (2006) showed that the power spectrum density of the babble noise obtained from Polish male speech has a distinctive increase in level in the region of frequencies higher than 5 kHz. It results from the fact that Polish has the greatest number of affricates among European languages. The high frequency maximum in the babble noise is something that was not observed in any of Byrne's spectra, and supports the need for this research. Additionally, although Byrne's results (Byrne *et al.*, 1994) became a landmark work for generations of researchers, it needs to be noted that these results were produced over 20 years ago. It cannot be ruled out that the recording conditions (different for almost every language group), technical capabilities of used devices or the averaging over the levels in frequency bands failed to reveal subtle differences between long-term spectra of the investigated languages, especially in the high frequency region.

The consideration for patient's language during fitting of a hearing aid can be observed in clinical practice of the recent years. NAL-NL2 fitting method can serve as an example, since its algorithm calculates the insertion gain in a different way depending on whether the patient speaks a tonal language or not. Such a language-oriented approach can be also found in research works of the last years (Chasin, 2008; 2011; Chasin, Hockley, 2013; Noh, Lee, 2012).

The above observations about ISTS design and the potential significance of individual language features have provided motivation for the current study. Its goal was to create a single-language counterpart of multi-language ISTS, to analyse and compare their acoustical characteristics. Polish language has been chosen because of the authors' origins and the fact that the Slavic-group of languages was neglected in the selection of ISTS' languages. The created signal is ISTS' counterpart in the sense that it has the same structure and fulfils all requirements of a test signal described in IEC 60118-15 (except for the language-related ones). This paper describes some aspects of the Polish Speech Test Signal's development procedure, presents the results of the comparative analysis of PSTS and ISTS, and discusses the results and their potential implications. This study is expected to contribute to the scientific pursuit that takes a language into consideration as a significant factor in the process of hearing aids fitting.

2. Development of PSTS as compared to ISTS

This section does not aim at describing in detail the development procedure of PSTS, which was meant to mirror the development procedure of ISTS described in detail by Holube *et al.* (2010). Instead, it focuses on the main differences between these two procedures.

2.1. The recorded text

The spoken materials recorded for ISTS were translations of the same content – a fairy tale – into each of the six languages. The benefit of such an approach consists in ensuring the same semantic load and level of vocabulary difficulty (in this case – basic vocabulary).

For the present study, it was essential to use a Polish text. It was decided that the text should consist of diversified and contemporary vocabulary referring to a variety of semantic domains, and not include specialist or professional terms. Moreover, the text recorded for the purpose of PSTS was phonemically balanced (Fig. 8b), which was not the case for ISTS. The phonemic balance means that the phonemic content of recorded text was closely akin to the phonemes' incidence in Polish speech. A comparison of phonemes' incidence in the recorded text and that obtained by Ziółko *et al.* (2009) on a large sample of Polish written texts is presented in Fig. 8b.

2.2. Recording conditions and preliminary processing

During four recording sessions, voices of 18 female speakers reading the text were recorded in the studio at the Institute of Acoustics, Adam Mickiewicz University. They were all native speakers, 21–60 years old, without speech impediments. The speakers were seated in front of a Neumann U87i directional microphone (located at a distance of 20–30 cm from the speaker's mouth), plugged into a YAMAHA DM1000 mixer. Monophonic soundtracks were recorded with 44 100 Hz sampling frequency and 24-bits resolution using a Samplitude Pro X digital audio workstation.

Preliminarily processing of recorded material included manual removal of unintended sounds using Samplitude software (e.g. mistakes and coughs) and automatic reduction of silent intervals to the maximum duration of 600 ms. The identification and reduction of silence was done automatically in Matlab, using an algorithm based on a power level threshold value, which was set to −27 dB relative to the peak value of the whole signal. This value was chosen empirically making sure (by means of careful auditory and visual inspection of the waveform) that the soft fragments of speech are not classified as silence and cut off. Choosing a higher threshold would result in trimming the speech signal in few cases of recordings, whereas a lower value would not remove nonspeech fragments that were contaminated with some low-amplitude background noise. Therefore, it was decided to apply the value of −27 dB as the most optimal in this case.

2.3. Selection of recordings

For ISTS the selection criteria of six recordings (out of the previously recorded set) included: the dialect, voice quality, the naturalness of pronunciation, and the median fundamental frequency. For PSTS the dialect criterion was not applicable, and the quality and naturalness were prerequisite at the initial stage of making the 18 recordings. The final criteria were the following two:

- proximity of their fundamental frequencies (median values),
- spectral affinity to the long-term average Polish female speech spectrum (PL-LTAFSS).

The selection was based on a compromise between them (see Table 1). Since no record of contemporary long-term average Polish spectrum of female speech (PL-LTAFSS) had been found, it was also created using the results of this study. A paper regarding the process of developing PL-LTAFSS is currently in preparation.

To assess the spectral affinity, 1/3-octave band levels (63–20 000 Hz) were calculated for each of the recordings, compared with 1/3-octave band levels of PL-LTAFSS and for each band the deviation was derived.

Table 1. Six selected recordings and their characteristics: median of fundamental frequency (column 1), fundamental frequency deviation (column 2), maximum difference $D_{\max}$ in 1/3-octave band levels between the recording and the reference band spectrum (column 3), frequency band in which $D_{\max}$ occurred (column 4).

	f_0 [Hz]	f_0 dev [Hz]	$D_{\max}$ [dB]	$f_{D\max}$ [Hz]
Recording 1	202	33	6.4	315
Recording 2	210	35	4.2	125
Recording 3	210	52	6.6	1600
Recording 4	213	49	3.4	125
Recording 5	221	33	8.9	125
Recording 6	206	36	6.5	4000

2.4. Reference LTASS and filtration process

Holube *et al.* (2010) have filtered their samples to the ILTASS of female speech described by Byrne *et al.* (1994) using FIR-filters between 100 and 16 000 Hz to improve the homogeneity of the speech material. It would not be in accordance with the aim of the study to make the spectral shape of Polish recordings resemble the international spectrum, therefore PL-LTAFSS was chosen as a reference. On the basis of differences between each recording's 1/3-octave band levels and the corresponding band levels of PL-LTAFSS, a set of FIR filters was designed in Matlab using frequency sampling method. After filtering, the deviation from PL-LTAFSS within the whole range of frequency bands did not exceed 3 dB.

2.5. Segmentation

In the next step, the recordings were split into short (100–600 ms long) segments. When determining the positions of segments' boundaries, it had to be ensured that:

- the boundaries would not break the continuity of a syllable,
- pauses equal to or longer than 100 ms, that naturally occur in the course of speech, would be placed at the end of a segment.

These requirements were the same as for ISTS. The difference consisted in the tools used to achieve the goal, chiefly because Polish text material was much longer than ISTS'.

In order to automate segmentation of PSTS source material, AnnotationPro (v.2.2.1.5) software was used (Klessa, 2015; Klessa *et al.*, 2013). Provided with an audio file and a phonetic transcription of recorded text, this software is able to split the utterance into syllables and set markers at their boundaries. As some of the syllable segments were shorter than 100 ms, they had to be manually merged with their neighbours. Whenever the splitting algorithm failed in its accuracy leaving an audible artefact at the end or beginning of a segment, the artefacts were eliminated by fine manual adjustments, supported by an aural assessment of isolated segments and visual control of the waveform. Localisation of speech pauses and shifting of segments boundaries (so that the pauses were located always at the end of segments), as well as the final splitting were done in Matlab. As a result of this process 7690 segments were obtained.

2.6. The order of speakers within a speech phrase

Both signals, i.e. ISTS and PSTS, were composed by drawing the segments from a pool and lining them up in a new order. However, to provide for key desired features of the signals, some restrictions were imposed on the randomness of the process. For ISTS, one of them was related to the changeability of languages from segment to segment and thus – to the changeability of speakers. The strategy of ISTS' development ensured that "*each language was selected randomly only once within each group of six consecutive segments*" (Holube *et al.*, 2010). PSTS does not have changeability of language, only of speaker. PSTS has in principle a fixed order of speakers and the order might be disturbed only at the junction of two neighbouring sections. It should be noted that both ISTS and PSTS signals are constructed of subsequent 15-, 10-, 10-, 10-, and 15-second long sections, which can be considered as independent, separate parts. A general structure of both signals is presented schematically in Fig. 1.

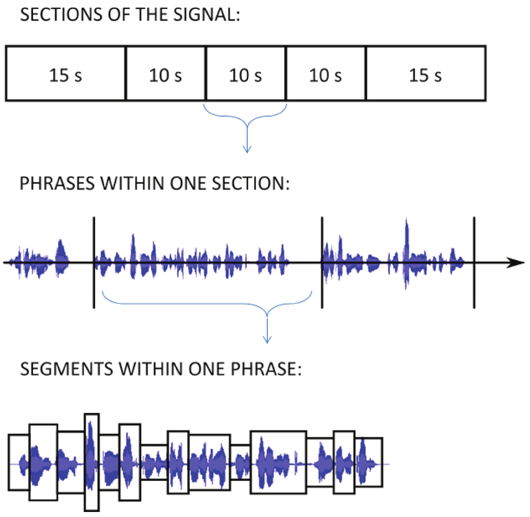

Fig. 1. A schematic diagram representing the structure of both signals.

2.7. Composition of sections and number of speakers

Automatic composition algorithm performed by Matlab ensured compliance with all segments composition requirements described by Holube *et al.* (2010). The sections generated by the algorithm consisted of speech phrases of different length. Their exact duration was set according to the probability calculated previously from the distribution of speech phrases durations in eighteen, previously made Polish speech recordings (not on the data used to generate ISTS).

Also in contrast to ISTS procedure, the preliminary auditory assessment was applied on-line during the process of creating PSTS sections. Just after a phrase was composed by the pseudo-random algorithm, it was played back and had to be accepted by the code executor (the first author of the article). Most of them were accepted, but the ones sounding grossly unnatural in the person's opinion were rejected. Unnaturalness could occur for example as a result of unmatched levels of neighbouring segments, strange, non-Polish-like intonation or accent distribution in a phrase. It was also desired to avoid meaningful concatenations of syllables. During this process, it was observed that all segments from one of the speakers (recording 6) stood out distinctly against all others. Every occurrence of this speaker's segment caused an impression of interruption and discontinuity. This effect was consulted and confirmed with numerous listeners and it was decided to eliminate this speaker' segments and repeat the procedure of generating sections with only five speakers.

2.8. Auditory assessment and selection of final sections

The final stage of signal composition was a selection of three 10-second long and two 15-second long sections. In contrast to the previous stage (auditory control of speech phrases' quality), the subjective auditory assessment was inherent in the process of final selection of five sections.

Ten listeners (not involved in the process of creating signal) participated in this study. Their task was to rate the naturalness of each section sound on a scale from 1 to 5 (with 1 being the worst and 5 the best grade, and with permissible grade step of 0.25). Every listener participated in two sessions, separated in time – one for the assessment of the 10-second sections, and the second one for the 15-second sections. The listeners were informed beforehand about the nature of the signals. They were forewarned not to expect a natural speech, but they were asked to assess the naturalness and homogeneity of the sound of every section. To avoid the effect of expectations-driven bias on the first judged sections, the whole set of sections was firstly presented to the listeners. After this preconditioning, the actual experiment was initiated. The sections were presented through a *Pioneer* SE-M390 headset, one after another, with 4-second long breaks in between for writing down its grade. Order of sections was the same for all listeners. By means of this procedure three highly rated shorter sections and two highly rated longer sections were selected and thus they formed the final PSTS.

Both development procedures (of ISTS and PSTS) used at a certain stage a subjective auditory assessment. In the case of ISTS it was realised after generating all sections and "*those which sounded most natural (especially at the beginning and the end) and exhibited the most homogeneous speech and pause distribution were selected*" (Holube *et al.*, 2010). By contrast, in the case of PSTS, rough, preliminary assessment was implemented already at the stage of generating individual speech phrases. Because of that, more effort was put into creating one section, but the resultant ones were marked by a higher quality and thus a smaller number of them was required. This also facilitated the stage of final selection which was performed by means of ten independent listeners' assessment and ranking.

3. Analyses and comparison between PSTS and ISTS

Comparative analysis of PSTS, ISTS and – if applicable – PL-LTAFSS was made, taking into account percentile analysis results, crest factors, long-term averaged spectra, spectrograms, duration of continuous speech fragments (phrases) and duration of pauses between them, modulation spectra, comodulation patterns, fundamental frequencies of signals, percentage of voiced and voicelessspeech fragments in the signals. Moreover, the phonetic content of PSTS was analysed.

3.1. Percentile analysis

Percentile analysis of both test signals was performed in 125 ms time windows and in 22 1/3-octave bands of central frequencies from 125 Hz to 16 000 Hz (IEC 60118-15, 2012). The acoustic pressure levels corresponding to the 30th percentile (soft speech), 65th percentile (moderately loud speech), 95th and 99th percentile (loud speech) were calculated. Results of the percentile analysis of PSTS and ISTS are presented in Fig. 2. In PSTS for the majority of bands the thirtieth percentile value was smaller than for ISTS (the difference amounting to 8 dB in the 500 Hz frequency band). The 99th percentile values were in general higher for PSTS, especially in the high frequency region (above 2.5 kHz) – amounting to 7 dB in 10 kHz frequency band. Also the difference between the 30th and 99th percentile was greater in PSTS than in ISTS. These values indicate a greater number of soft speech fragments and greater dynamic range of PSTS. For comparison purposes the differences between the 30th and 99th percentiles of ISTS and PSTS signals were compared with those given by Cox *et al.* (1988).

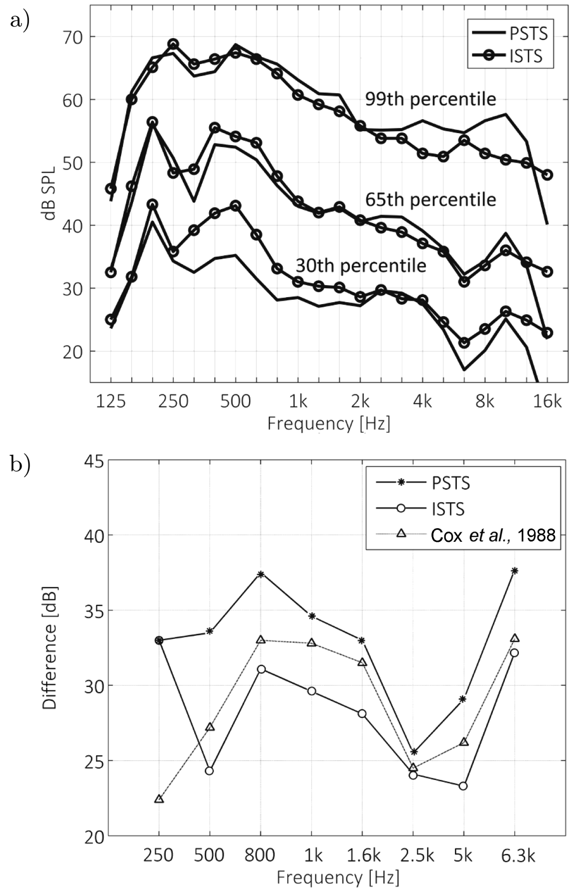

Fig. 2. a) Distribution of short time sound pressure level (SPL) values in subsequent 1/3-octave bands expressed in terms of percentile values; b) comparison of differences between the 99th and 30th percentile values for PSTS and ISTS and the values given by Cox *et al.* (1988).

However, it should be noted that Cox *et al.* (1988) analysed slightly shorter speech signals of American English, in a smaller number of 1/3-octave bands, so this comparison is more qualitative than quantitative. The differences between the 30th and 99th percentiles show similar tendencies of increase or decrease in subsequent frequency bands (Fig. 2b). According to the results, the dynamic range observed in 69% of the loudest instantaneous values of the signal was greater in PSTS than in ISTS and also greater than that calculated by Cox *et al.* (1988) for the USA English speech.

3.2. Crest factor

Crest factor (CF) is defined as a ratio of the peak value of a signal to its effective (root mean square, RMS) value. The CF values for PSTS and ISTS were 13.7 dB and 9.7 dB respectively. According to Chasin (2008) the typical value of this parameter is 12 dB. The difference in CF value, similarly as the results of percentile analysis imply that the dynamic range of PSTS is greater than that of ISTS.

3.3. Long-term averaged spectrum

The long-term averaged spectra of PSTS and ISTS were determined for 1/3-octave bands using ArtemiS software (HeadAcoustics). The results are presented in Fig. 3, also showing the Polish long-term averaged spectrum of female speech, PL-LTAFSS. The curves obtained for PSTS and PL-LTAFSS are similar and exhibit local maxima for the frequencies 3–4 kHz and 8–10 kHz, which do not appear in the ISTS spectrum.

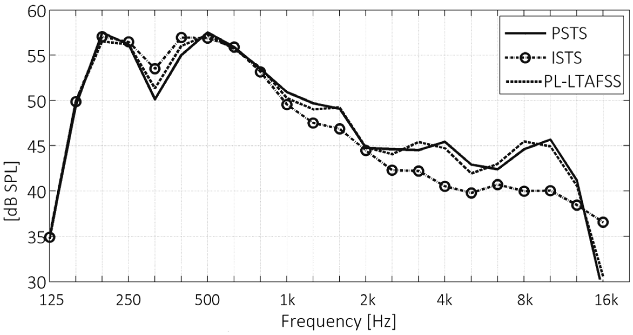

Fig. 3. Long-term averaged spectrum of PSTS, ISTS, PL-LTAFSS (Habasińska, 2015). All spectra were normalized to have a total sound pressure level of 65 dB SPL.

3.4. Spectrograms

1/3-octave spectrograms for PSTS and ISTS were calculated in ArtemiS for 10 second signal sections(see Fig. 4). In the PSTS spectrogram the pauses in speech are more pronounced than in the one of ISTS. This is most certainly a consequence of removing the parts of the waveform in which the speakers took breaths. In both spectrograms typical speech-like formant structures and fundamental frequency contour can be observed.

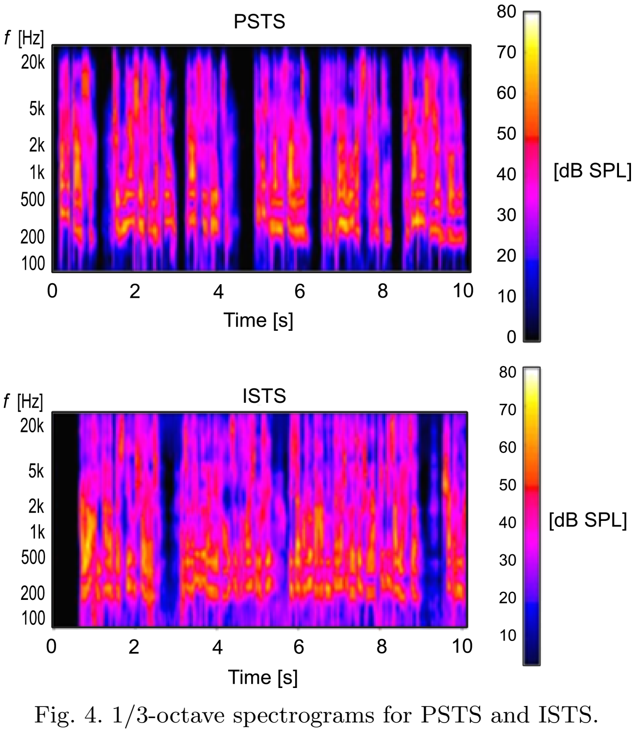

Fig. 4. 1/3-octave spectrograms for PSTS and ISTS.

3.5. Duration of continuous speech fragments

The distribution of durations of continuous speech fragments (called also phrases), defined as a timespan between two over 100 ms long pauses, was analysed in PSTS. Histogram presenting the relative frequency of speech phrases of a given duration is shown in Fig. 5. The most frequent durations of speech phrases in PSTS are those of 600–800 ms. The histogram was fitted with the Weibull function defined by the following equation (1), where $\alpha = 1.6$, $\beta = 4.82$ and the coefficient of determination was 0.8456

$$f(x) = \begin{cases} 0, & x < 0, \\ \dfrac{\alpha}{\beta}\left(\dfrac{x}{\beta}\right)\exp\left(-\left(\dfrac{x}{\beta}\right)^{\alpha}\right), & x \geq 0, \end{cases} \tag{1}$$

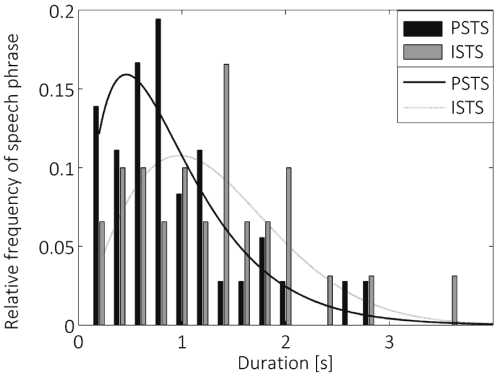

Fig. 5. Histogram of continuous speech fragments duration in PSTS and ISTS.

Figure 5 presents also the histogram of durations of speech phrases for ISTS – the values are taken from Holube *et al.* (2010). Note that in ISTS the most frequent speech signal duration (1.6 s) is almost two times longer than in PSTS.

3.6. Duration of pauses

For the purpose of this analysis, pause was defined as a period of silence lasting for at least 10 ms. The median of pause durations in PSTS was 80 ms. Wilson *et al.* (1983) have shown that the mean duration of pauses in normal fast speech is 42–49 ms, while in slow speech it is 130 ms. In PSTS and ISTS the most frequent were short pauses lasting less than 75 ms.

3.7. Comodulation patterns

Comodulation pattern refers to a graphical representation of the matrices of correlation coefficients between temporal changes of the envelope of a given band and every other frequency band in the same signal. Frequency bands used for this analysis have a width of 1/3-octave. The patterns obtained for both signals (Fig. 6) resemble typical speech-like patterns and are dissimilar to the ones found in signals of different class, i.e. ICRA-5 or white noise (Habasińska, Skrodzka, 2017). However, the patterns for ISTS and PSTS show differences in the following pairs of frequency bands: 315/630 Hz (which is highly correlated in PSTS but not in ISTS) and 4000/5000 Hz (where the contrary is the case). In a rough approximation, PSTS displays higher correlation for the pairs of frequency bands above 8 kHz, below 250 Hz and in the region between 315 and 1600 Hz, while in the other regions higher correlation is observed in ISTS (Holube *et al.*, 2010).

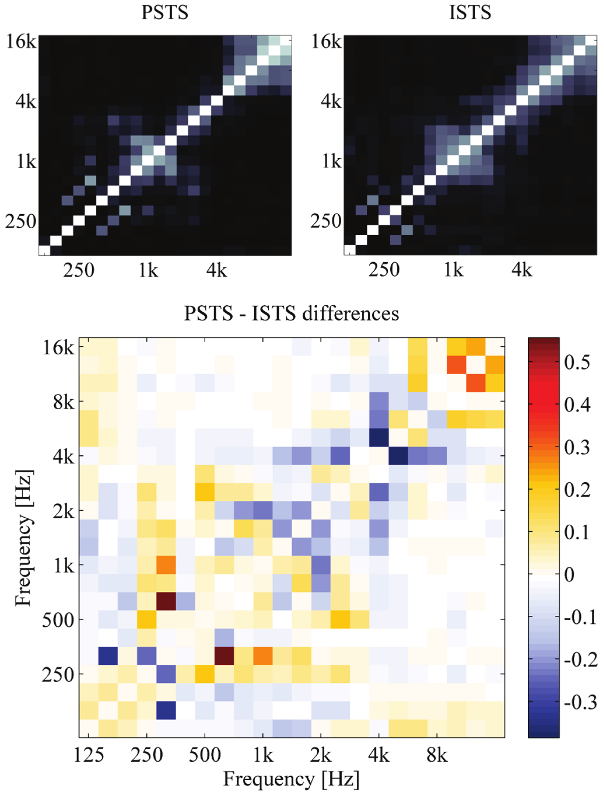

Fig. 6. Comodulation pattern for PSTS (upper left), ISTS (upper right) and a matrix containing differences between correlation coefficients of PSTS and ISTS (bottom) (Habasińska, Skrodzka, 2017).

3.8. Percentage of voiced and voiceless speech fragments

The percentages of voiceless speech fragments in PSTS and ISTS (Fig. 7) were calculated using Praat (Boersma, Weenink, 2015) and reached 55.5% in PSTS and 43.1% in ISTS.

The pauses present in the signals might be treated by the algorithm as voiceless speech fragments, therefore to eliminate the impact of differences in the pause duration in ISTS and PSTS, these signals were subjected to repeat analysis after limiting the pauses duration to 10 ms. As expected, the calculated percentage of voiceless speech fragments in these two signals was decreased significantly. However, the difference in the percentage contribution of voiceless speech fragments between the two signals increased (38.5% in PSTS and 19.23% in ISTS; Fig. 7). Thus, the percentage contribution of the voiceless speech fragments in the Polish signal is two times higher than in the international signal ISTS.

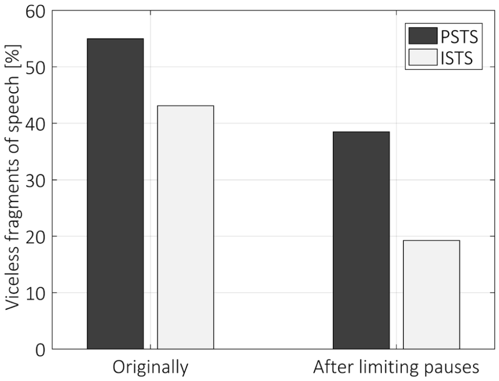

Fig. 7. Percentage of voiceless fragments of speech in PSTS and ISTS – before and after limiting pauses duration.

3.9. Fundamental frequency of the signal

The fundamental frequency of PSTS, its mean value, median and standard deviations were calculated using the Praat program configured for analysis of female voices (Podesva, Sharma, 2013), and corresponding values for ISTS were taken from Holube *et al.* (2010). The comparison of the values related to PSTS, ISTS and particular recordings are given in Table 2. The fundamental frequency of PSTS signal (mean value and median) is by 15 Hz higher than that of ISTS.

3.10. Phonemic content of PSTS

The phonemic content of PSTS was logged during the signal generation procedure and statistically analysed afterwards. The results were compared with the Polish phoneme statistics obtained on a large set of written texts by Ziółko *et al.* (2009) (Fig. 8b).

The differences between the frequency of phoneme occurrence in PSTS and in the reference data provided by Ziółko *et al.* (2009) are bigger than between the text recorded and the reference (Figs. 8a and 7b). However, it should be noticed that the shorter the text, the more difficult it is to achieve the phonemic balance. The text material of a one minute long PSTS is composed of only 649 phonemes, grouped in 250 segments. The choice of the segments was determined by different factors, e.g. speaker, position within a speech

Table 2. Mean values, median and standard deviation of the fundamental frequency calculated for particular speech recordings and for PSTS (original results) and ISTS (from Holube *et al.* (2010)).

	PSTS/Polish Recordings			ISTS/		
	Median f_0 [Hz]	Mean f_0 [Hz]	σ [Hz]	Median f_0 [Hz]	Mean f_0 [Hz]	σ [Hz]
Speech-like test signal	211	216	37.5	195	196	43.0
All speakers	209	213	32.2	203	207	44.3
Speaker 1	201	204	24.8	194	201	45.1
Speaker 2	209	212	27.5	201	206	45.0
Speaker 3	209	217	41.4	204	208	54.6
Speaker 4	209	212	38.0	205	205	36.4
Speaker 5	219	220	29.2	207	209	34.8
Speaker 6	–	–	–	208	210	49.7

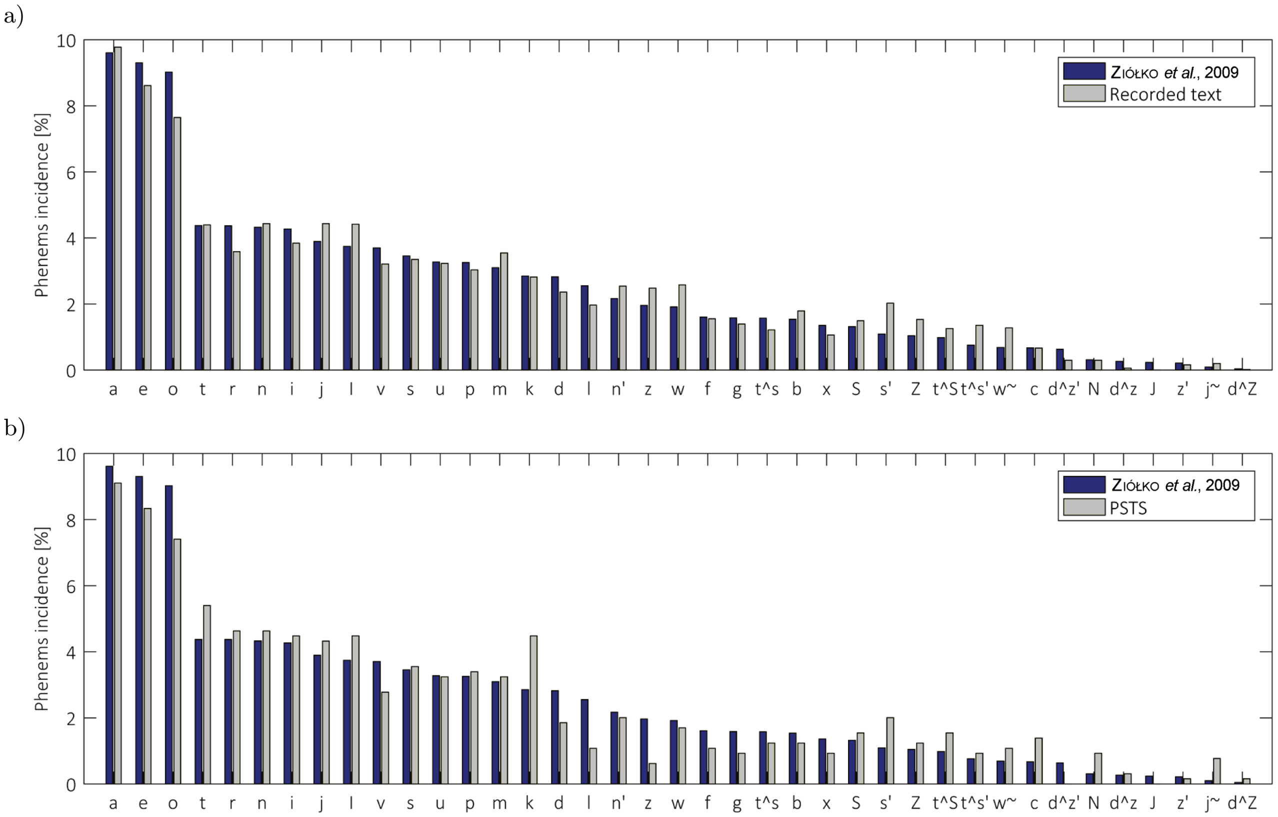

Fig. 8. a) Percentage content of individual phonemes in the recorded text compared to phonemes' incidence model established for Polish speech; b) percentage content of individual phonemes in PSTS and given by Ziółko *et al.* (2009).

phrase, duration, fundamental frequency. As one of the aspects of the analysis it is worth noting that in PSTS the voiceless phonemes make 27% of all phonemes (in the text recorded by the female speakers the percentage contribution of voiceless phonemes was 22.3%).

4. Conclusions

The results obtained and the comparative analysis of ISTS and PSTS permit drawing the following conclusions.

- The long-term averaged PSTS spectrum in the frequency ranges 3–4 and 8–10 kHz shows two characteristic maxima, while the ISTS spectrum falls smoothly in the corresponding frequency region.
- According to the percentile analysis, PSTS contains a greater number of soft speech fragments than ISTS (the difference in 30th percentile values amounting to 8 dB in the 500 Hz frequency band), and – in the high-frequency region – greater number of loud speech fragments. This is related to a relatively bigger variety of instantaneous SPL values.
- The crest factor (CF) for PSTS is by 4 dB higher than for ISTS. Wider dynamic range of PSTS can

be observed also in the results of percentile analysis.

- Having limited the pause duration in both signals to 10 ms, two times higher percentage content of voiceless speech fragments was obtained in PSTS than in ISTS. It is not a surprising result because one of characteristic features of Polish language is a high content of fricatives and affricates sounds.
- The pauses in PSTS are more frequent and statistically longer as evidenced in the spectrograms. Probably it is a consequence of a fluency of reading by Polish female speakers – they read the text slower than the speakers recording the texts for ISTS.
- In ISTS the most frequent duration of continuous speech fragments (1.6 s) is twice as long as in PSTS of 0.6–0.8 s. The durations of continuous speech fragments in PSTS are a bit shorter than in natural speech, but their relative distribution resembles that obtained from live-speech recordings.
- The mean value and median of the fundamental frequency in PSTS are by 15 Hz higher than in ISTS.

Similarities between PSTS and ISTS are revealed in their comodulation patterns and spectrograms, which exhibit prominent contours of the fundamental frequency and formant transitions. The fact that the above mentioned features are similar to the analogous features of live speech indicates that both signals have preserved the most important characteristics of live speech. The differences between PSTS and ISTS reveal the influence of the language and point out the specificity of Polish language.

The objective of the next planned research is to verify whether the differences between ISTS and PSTS are significant enough to affect the effectiveness of hearing aid fitting for Polish language speakers.

The problem considered in this study applies not only to Polish language but also to many other languages not used for ISTS generation. Analogous studies for other languages could lead to development of signals based on a given specific language or group of languages, which would serve as local standards in hearing aids fitting practice. To summarise, in the field of hearing aid measurements, the need for one universal standard signal has been fulfilled by ISTS, but in the field of hearing aid's fitting, the signal that is used should permit optimum intelligibility and thus communication in the language that is most often used by the aided person.

Acknowledgments

The authors wish to thank Katarzyna Klessa, who contributed greatly to this work by introducing AnnotationPro software and by providing technical support and guidance during the process of segmentation.

References

1. Boersma P., Weenink D. (2015), *Praat: doing phonetics by computer* (Computer program), version 5.4.09, retrieved May 11, 2015, from http://www.praat.org.
2. Byrne D. *et al.* (1994), *An international comparison of long-term average speech spectra*, Journal of Acoustical Society of America, **96**, 4, 2108–2120.
3. Chasin M. (2008), *How hearing aids may be set for different languages*, Hearing Review, **15**, 16–20.
4. Chasin M. (2011), *Setting hearing aids differently for different languages*, Seminars in Hearing, **32**, 182–188.
5. Chasin M., Hockley N.S. (2013), *An automated system to improve hearing aid settings for non-English speakers*, Hearing Review, **20**, 4, 28–32.
6. Cox R.M., Matesich J.S., Moore J.N. (1988), *Distribution of short-term rms levels in conversational speech*, Journal of Acoustical Society of America, **84**, 3, 1100–1104.
7. de Boysson-Bardies B., Sagart L., Halle P., Durand C. (1986), *Acoustic investigation of cross linguistic variability in babbling*, Precursors of Early Speech, **16**, 1, 113–126.
8. Dworsack-Dodge M., Switalski W. (2012), *Current practices in modern probe microphone measurement*, URL: http://www.otometrics.com.
9. Habasińska D. (2015), *The development of the Polish speech test signal for measuring and fitting hearing aids* [in Polish: *Stworzenie polskiego testowego sygnału mowopodobnego do wykorzystania w miernictwie i dopasowaniu aparatów słuchowych*, Master Thesis, Adam Mickiewcz University Poznań.
10. Habasińska D., Skrodzka E. (2017), *Development and analysis of the Polish Speech Test Signal in view of the IEC 60118-15 Standard*, DAGA 2017, 127–128.
11. Holube I., Fredelake S., Vlaming M., Kollmeier B. (2010), *Development and analysis of an International Speech Test Signal (ISTS)*, International Journal of Audiology, **49**, 12, 891–903.
12. International Electrotechnical Commission (2012), IEC 60118-15, *Electroacoustics – Hearing aids – Part 15: Methods for characterizing signal processing in hearing aids with a speech-like signal.*
13. Jassem W. (2003), *Polish*, Journal of the International Phonetic Association, **33**, 1, 103–107.
14. Klessa K., (2015), *Annotation Pro* [software], Katarzyna Klessa.
15. Klessa K., Karpiński M., Wagner A. (2013), *Annotation Pro-a new software tool for annotation of linguistic and paralinguistic features*, Proceedings

of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop, Aix en Provence, pp. 51–54.

16. Kossek P., Dworsack-Dodge M. (2010), *Dynamic REM with Percentile analysis*, www.otometrics.com.

17. Noh H., Lee D.H. (2012), *Cross-language identification of long-term average speech spectra in Korean and English: toward a better understanding of the quantitative difference between two languages*, Ear Hear, **33**, 3, 441–443.

18. Ozimek E., Kutzner D., Sęk A., Wicher A., Szczepaniak O. (2006), *The Polish sentence test for speech intelligibility evaluations*, Archives of Acoustics, **31**, 4S, 431–438.

19. Podesva R.J., Sharma D. [Eds.] (2013), *Research Methods in Linguistics*, Cambridge University Press, Cambridge, pp. 375–396.

20. Wilson R.H., Margolis R.H. (1983), *Measurement of auditory thresholds for speech stimuli*, [in:] Konkle D.F., Rintelmann W.F. [Eds.], *Principles of speech audiometry*, Academic Press, Baltimore, pp. 79–126.

21. Ziółko B., Gałka J., Ziółko M. (2009), *Polish phoneme statistics obtained on large set of written texts*, Computer Science, **10**, 97–106.