Short description of the International Speech Test Signal (ISTS)

Inga Holube[1], Marcel Vlaming[2], EHIMA-ISMADHA Working Group[2]

[1] *Center of Competence HörTech and Institute of Hearing Technology and Audiology, Oldenburg, Germany, Email: Inga.Holube@fh-oow.de;*

[2] *EHIMA-ISMADHA WG, Email: mv@ehima.com*

Summary

For analyzing the processing of speech by a hearing aid, a standard test signal is necessary which allows for reproducible measurement conditions and which features all or most relevant properties of natural speech. Those properties are e.g. the modulation spectrum and the fundamental frequency as well as its harmonics. Existing artificial signals fulfill these requirements inadequately and recordings from natural speakers represent only one language and are therefore not internationally applicable. Therefore, the European Hearing Instrument Manufacturing Association (EHIMA) has set up the ISMADHA working group and has initiated this project resulting in an International Speech Test Signal (ISTS). The ISTS is based on natural recordings but is largely non-intelligible because of segmentation and remixing. The intention is to include this test signal with a new measurement method for a new hearing aid standard.

A full description of the development and analysis of the ISTS is found in [1].

Speech recordings

21 female speakers in six different mother tongues (American English, Arabic, Chinese, French, German and Spanish) were reading the story "The north wind and the sun" [2] several times using natural articulation. The recordings were done with a Neumann KM184 directional microphone and sampled with a sampling frequency of 44.1 kHz and a resolution of 24 bit in a modified office space (reverberation time of 0.5 s at 500 Hz). For each language, one recording of one speaker was selected. Selection criteria were the regional provenance of the speakers, the voice quality (e.g. croakiness) and the median fundamental frequency. The recorded speech material was filtered to the International Long Term Average Speech Spectrum of female speech between 100 Hz and 16 kHz according to [3] such to optimize the homogeneity of the speech material. In addition, the distribution of the speech duration between longer speech pauses (above 100 ms) was compiled and a probability function was fitted to this distribution as needed for the mixing of the recordings. In this distribution function the duration of the speech pauses was limited to 650 ms.

Segmentation of recordings

The recordings were fractionized in segments using an automatic procedure: Initial segments with a duration of 500 ms were taken from the recordings. From these 500 ms segments, the power was analyzed in 10ms-intervals for the last 400 ms. From that the 10 ms-interval with the lowest power was selected. Within that interval the lowest absolute value was picked. The resulting segment then contained the recording from the start of the initial 500ms-segment until this lowest absolute value. The next 500ms-segment started directly after this lowest absolute value. This automatic segmentation had to be modified by hand to avoid cutting points within vowels and associated phonemes as much as possible. The resulting segments had a duration between 100 and 600 ms. Speech pauses with a duration of more than 100 ms were kept within the same segment as the previous speech utterance to ensure their natural position. Those segments including long pauses as well as the following "begin-segments" were marked.

Mixture of segments

The segments were attached to each other in random order to generate sections with a duration of 10 s and 15 s. During this procedure, the segments were modified with a Hanning window with a shoulder of 1 ms on each end to avoid audible artifacts. In addition, the language was changed from segment to segment and each language was selected once within six consecutive segments. Each segment was used once within a 10 s- or 15 s-section. To minimize the steps of the fundamental frequency, the fundamental frequency was analyzed within the first and the last 50 ms of each segment. When two voiced segments were attached to each other, only changes of the fundamental frequency up to 10 Hz were allowed. If this criterion was violated, another segment was selected. The combination of a voiced and an unvoiced as well as two unvoiced articulations were always possible. Those segments with pause durations of more than 100 ms were selected when the speech duration was exceeding a value calculated based on the probability distribution described above. This limitation guarantees a natural distance between the speech pauses. After each speech pause, a "begin-segment" was selected from a different language. At the end of each 10 s- and 15 s-section, a segment including a speech pause was selected and limited to the necessary duration of each section. All generated sections were filtered again to the international female spectrum described in [3]. The ISTS with a duration of 60 s was composed from the 10 s- and 15 s-sections. Other durations in steps of 5 ms (without 5 ms and 55 ms) are possible. For hearing aid measurements, a duration of 15 s should be used to allow the signal processing algorithms to adjust to the signal. Thereafter, a measurement duration of 45 s should be used. To allow for a rough estimation of the measurement results, it should be possible to limit the measurement duration to 10 s.

Analysis of test signal

The ISTS composed by the procedure as described above was analyzed in respect to different criteria and compared to

the original recordings. It was shown that the ISTS agrees to natural speech in all relevant criteria. The most important results for the ISTS with a duration of 45 s is summarized below.

Long-term spectra

The long-term spectra of the ISTS as well as the 10 s- and 15 s-sections deviate by less than 1 dB from the international long-term female speech spectrum of [3].

Short-time spectra

The short-time spectrum of the ISTS shows steps in the fundamental frequency at several degrees as can be observed also in the original recordings for the different languages.

Fundamental frequency

The median of the fundamental frequency of the ISTS is 196 Hz, compared to a median of 203 Hz for the speakers in the original recordings. This is regarded as sufficiently similar. The standard deviation is 44 Hz for the ISTS which is the same as for the original recordings.

Modulation spectra

The modulation spectra of the ISTS as well as for the original recordings filtered in 1/3-octave bands show a maximum in the range of 2-8 Hz. Systematic deviations were not observed.

Comodulation analysis

The comodulations were analyzed by correlating the envelopes of the signal filtered in 1/3-octave bands. The strength of the cross correlation is reduced with increasing distance between the 1/3-octave bands. This applies for the ISTS as well as for the original recordings.

Pause duration

The distributions of the speech pauses and their duration correspond to the original recordings. However, the shorter duration of the ISTS results in a slight more unevenly spreading compared to the original recordings. The ratio of pause duration versus signal duration is 1 over 6.

Percentile distribution

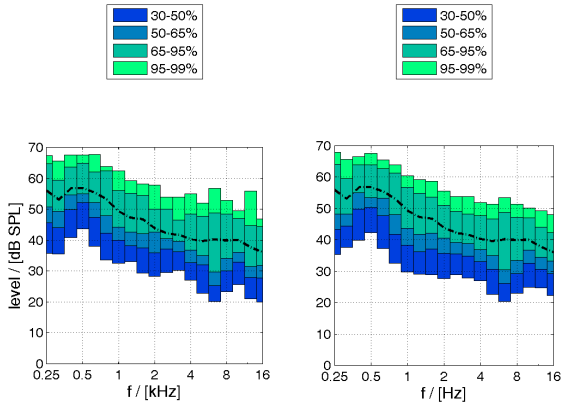

Figure 1: Percentile distribution of the levels in 125 ms-windows for a 10 s-section (left) and a 45 s-section (right) of the ISTS together with the international speech spectrum taken from [2] (dashed line).

The signals were filtered in 1/3-octave bands and the levels were calculated in 125 ms-windows (50% overlap). From this level distribution, the differences between the 99th , 65th and the 30th percentiles were calculated. The differences between the 99 and 30 percentiles are between 20 and 30 dB for the ISTS as well as for the original recordings (see Fig. 1).

Fraction of voiceless fragments

The fraction of voiceless fragments is 44% for the ISTS and is therefore slightly above the average value of 35% for the original speech recordings.

Instantaneous amplitude distribution

The distribution of the instantaneous amplitudes of the ISTS is very similar to that of the original speech recordings.

Crest-Factor

The CREST-factor of the ISTS has a value of 17 and is therefore very similar to the value of 18 for the original speech recordings.

Acknowledgement

Thanks to

- The ISMADHA working group for the very fruitful and intense cooperation and expert monitoring: Marcel Vlaming, VUmc-Amsterdam, Nicolai Bisgaard, GN Resound, Brian Pedersen, GN Resound, Volker Kuehnel, Phonak, Frank Rosenberger, Siemens, Ivo Merks, Starkey Laboratories, Carsten Paludan Müller, Widex, Johnny Andersen, Oticon, Todd Fortune, Interton.
- The European Hearing Industry Manufacturers Association, EHIMA, for initiative and financial support.
- Harvey Dillon, NAL, Kitte Geidser, NAL, Ake Olofsson, Karolinska Institutet, Robyn Cox, University of Memphis for advising on the specification and development of the signal.
- Birger Kollmeier, Volker Hohmann, Jörg Bitzer, Kirsten Wagener and Stefan Fredelake for fruitful discussions and support.
- Richard Schultz-Amling and Stefan Fredelake for coding the analysis methods and segmentation procedure in Matlab and trying to find the optimal solution, Jörg Bitzer for the filter bank and Jörn Anemüller for the comodulation analysis.
- Monika Kappelmann for organizing the speakers, Marco Wilmes and Jan Schaffmeister for doing the recordings and Björn Ohl and Marco Wilmes for measuring the reverberation time of the studio and the frequency response of the microphones.

References

[1] Holube I, Fredelake S, Vlaming M, Kollmeier B, (2010): Development and Analysis of an International Speech

Test Signal (ISTS), International Journal of Audiology, 2010; 49: 891–903.

[2]Handbook of the International Phonetic Association, Cambridge University Press.

[3]Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wibraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C., Kiessling, J., Kotby, M., Nasser, N., El Kholy, W., Nakanishi, Y., Oyer, H., Powell, R., Stephens, D., Meredith, R., Sirimanna, T., Tavatkiladze, G., Frolenkov, G., Westerman, S. und Ludvigsen, C.: An international comparison of long-term average speech spectra. J. Acoust. Soc. Am. 96 (1994), 2108-2120.