ARCHIVES OF ACOUSTICS
31, 4 (Supplement), 431–438 (2006)

THE POLISH SENTENCE TEST FOR SPEECH INTELLIGIBILITY EVALUATIONS

E. OZIMEK, D. KUTZNER, A. SĘK
A. WICHER, O. SZCZEPANIAK

Adam Mickiewicz University
Institute of Acoustics
Department of Room Acoustics and Psychoacoustics
Umultowska 85, 60-614 Poznań, Poland
e-mail: ozimaku@amu.edu.pl

(*received June 15, 2006; accepted September 30, 2006*)

The main purpose of this study was to prepare Polish sentence tests for measuring speech intelligibility against an interfering noise. The tests were developed from meaningful sentences taken from everyday utterances, TV, literature etc. Two sets of sentence lists were created. The first set was optimized for the so-called binary scoring and was composed of 25 lists of 20 sentences each, while the second set was prepared for word-based scoring and was composed of 22 lists of 20 sentences each. The respective lists were statistically and phonemically balanced, i.e. they produced comparable psychometric functions and revealed comparable phonemic distribution. The mean SRT (Speech Reception Threshold) and S_{50} (slope of a psychometric function at the SRT point) were: -6.1 dB and 29.5%/dB for the binary scoring and -7.4 dB and 26.7%/dB for the word-based scoring, respectively. The test lists comply with the requirements of the high quality test for measuring speech intelligibility of the Polish language.

Key words: speech intelligibility, sentence tests, psychometric functions, masking.

1. Introduction

In literature a few sentence tests have been described for measuring speech reception threshold (SRT – signal-to-noise ratio that yields 50% speech intelligibility) against noise [4–6, 8]. Some authors have shown that SRT depends only on the signal-to-noise ratio (SNR) [11] but according to others SRT depends both on SNR and on the presentation level [3, 13]. So far there is no Polish sentence test for measuring speech intelligibility in interfering noise. There are only a few Polish word tests for measurement the speech intelligibility. One of them worked out by PRUSZEWICZ *et al.* [9, 10] consists of different articulation lists. Each list contains 20 words from among the most frequent monosyllabic Polish nouns. The so-called Corpora test constitutes a collection of recordings of 114 short sentences and 20 numbers in each set, pronounced by 70 speakers [2]. It has been used mainly in the study of automatic recognition of speech.

Another test is a collection of 20 logatome sets, each set composed of three lists and each list containing 100 logatomes [1]. This test has been mainly used for assessment of the transmission quality of electroacoustic and teletransmission systems.

Sentence tests seem to be much more appropriate materials for speech intelligibility measurements since they reflect a real speech and produce much steeper psychometric function than the word or logatome tests. Due to high reliability and efficiency the sentence tests are often used in experiments concerning evaluation of various algorithms implemented in hearing aids [16] and spatial hearing in masking conditions [7].

It is known from literature that speech material used for measuring intelligibility should produce very steep psychometric functions so as to be able to detect changes in intelligibility at small differences of signal to noise ratio [5, 8, 14, 15]. To keep high accuracy of measurement, the intelligibility across different lists should not vary significantly. Moreover, the test lists should show high intra-comparability and low intra-variability, i.e. they should produce similar results.

The present study deals with the preparation and evaluation of the new Polish sentence test to be used in speech reception threshold measurements in noise. The test is structurally similar to the Dutch tests [8, 11, 14], the American test [6] and the German test [5].

2. Preparation of speech material and recordings

The test was prepared at two stages. At the first stage, about 3500 sentences were selected automatically from a large database containing about 16 millions of sentences taken from everyday speech, literature, TV, theatre etc., available in a digitized format. All of them were consistent with a fundamental definition of "sentence", i.e. they were compound of a subject and an object and contained normal everyday contexts. The following criteria were used in the automatic selection of the sentences [14]: the total number of syllables in a sentence should be equal to eight or nine; the words in sentences should not contain more than three syllables; the sentences should not contain punctuation characters and capitals (excluding the initial capital). Sentences chosen in this way were different, i.e. no duplicate sentences were selected. The second stage of the sentence selection was done manually on the basis of the following criteria [14]: the sentence tests should fulfill grammatical and syntactical correctness rules and semantic neutrality, which excluded political, war or sex topics, for example. Questions, proverbs, proper names and exclamations were eliminated. This process reduced the set of sentences to 1200. The chosen sentences were read out in a recording studio by a male professional speaker in a natural intonation, keeping approximately the same loudness level in time. Recording was performed using the capacity microphone Neumann U87 which functioned in an omni-directional mode so as to eliminate the so-called proximity effect leading to amplification of low-frequency spectral components of a recorded signal. The microphone output fed one of the input channels of the Yamaha 02R mixer. In the mixer the microphone signal was pre-amplified and converted into digital domain at the sampling rate of 44.1 kHz and with resolution of 24 bits and then high-pass filtered at a cut-off frequency of 80 Hz. After the processing signals were sent (an optical connection ADAT-type) to a PC and were stored on a computer hard disc using Samplitude Pro v.8.2 software.

3. Interfering noise

The masking noise was generated by means of summing up all the recorded sentences and normalizing rms value of a resultant wave, creating the so-called speech babble noise. All waveforms were shifted with respect to each other in the time domain, and, additionally, some of them were reversed in the time domain. The shift magnitudes and indexes of those signals were random. As a result, a 15-sec realisation of the speech babble noise was obtained. The main advantage of such masker is that for a given speaker average SNRs in the respective frequency bands (auditory filters) are kept constant during the masking measurements. Figure 1 depicts power spectrum density of the babble noise masker.

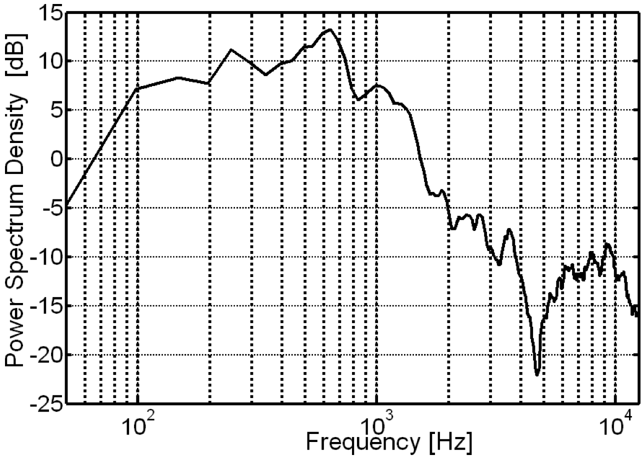

Fig. 1. Power spectrum density of interfering noise used in measurements.

4. Apparatus and measurement method

During the measurements, the recorded sentences were mixed digitally with the masking noise and presented to the subjects at five signal-to-noise ratios (SNR) varied from -9 to -1 dB. The noise sound pressure level value was normalized to 70 dB SPL, thus SNR depended on the sound pressure level of the sentence waveform. The signals were played back by the Tucker Davies Technology (TDT) System III with the 24-bit digital real-time signal processor RP2 and the headphone amplifier HB7 and were presented monaurally via the Sennheiser HD 580 headphones. The measurements were controlled by a software written in *Matlab 6.5* (*MathWorks*).

35 normally-hearing subjects participated in the experiments and a given sentence was presented to a respective subject only once. The subject's task was to repeat a presented utterance as precisely as possible. Two independent data collecting strategies were employed: "typed response", i.e. on a keyboard and "oral response" registered by the Sennheiser E914 microphone, pre-processed in the Yamaha MG10/2 and stored on a PC hard disc as a wav. file. Possession of the oral responses turned out to be crucial when the listeners made some typing mistakes.

5. Data analysis and derivation of psychometric functions

Estimation of the intelligibility score of the sentence material was made by two ways. In the first way, namely in the binary scoring method [14], the subject's response was treated as correct when all the sentence words were repeated correctly. In this case the score was 100%, otherwise i.e. in case of any error the score was set to 0%. The second way of the sentence intelligibility estimation was based on the correct repetition of the consecutive words in the sentence. The intelligibility score was determined as a ratio of the number of correctly repeated words to the total number of words in the sentence, multiplied by 100% [5]. It should be stressed that the employed scoring methods have some advantages and disadvantages. For example, the binary scoring is more simpler and faster than the word-based scoring, however it might be inapplicable for the subjects with profound hearing loss since they would never repeat the whole utterance correctly. In accordance, two list sets optimized for the binary and the word-scoring method were decided to be composed.

Some part of the response analysis was made automatically and some part had to be done manually. As a result, 25200 data related to the binary scoring and 116634 data for word-based scoring were obtained. Subsequently, for a given sentence mean intelligibility scores were determined at different SNRs by means of averaging the intelligibility data across subjects. Finally, psychometric function were fitted to these data by means of least-mean-square (LMS) method. The psychometric function was described by the standardized cumulative normal distribution (1):

$$\varphi(\mathrm{SNR}) = \frac{100}{\sqrt{2\pi}} \int\limits_{-\infty}^{(\mathrm{SNR}-\mathrm{SRT})/\sigma} e^{-t^2/2}\,\mathrm{d}t. \tag{1}$$

The function $\varphi(\mathrm{SNR})$ contains two parameters: SRT (the signal-to-noise ratio that produces 50% correct responses) and σ (the standard deviation which reflects scattering of the data). There is a direct relationship between the slope (S_{50}) of such a psychometric function and the standard deviation. The relationship is given by the expression (2):

$$S_{50} = \frac{100}{\sigma\sqrt{2\pi}}\,. \tag{2}$$

In order to compose highly reliable and accurate sentence materials, the psychometric functions corresponding to the sentences tested should show high values of S_{50}

and minimal spread of SRT values with respect to mean SRT. According to these requirements, 500 sentences for the binary scoring method and 440 sentences for the word-based scoring method were selected.

6. Composition of final sentence list sets

The last stage of developing the Polish sentence materials was a composition of the final list sets which should fulfill the following criterions:

- the lists should be statistically equivalent, i.e. an average SRT and S_{50} of the lists must not depend on the list index,
- the lists should contain a phonemically comparable linguistic content.

A special algorithm was prepared and implemented in Matlab 7.0 (MathWorks) which realized some Monte Carlo simulations. The composition steps were as follows:

- a random permutations of 500 and 440 sentences for the binary and the word-based scoring, respectively, were created,
- the random series of the sentences were grouped in 25 lists and 22 lists of 20 sentences each,
- in the last step the sentences (in written form) were translated into the phonemic code (SAMPA-broad) taking into account the so-called co-articulation effects using phoneme distribution rules [12].

If the sentence permutation meeting the above criterions was found, a new permutation was generated and the above mentioned steps were repeated, whereas the range related to the phonemic balance was narrowed by a constant value of $\pm 0.05\%$. This algorithm led to a composition of statistically independent list sets, in which the range for respective phonemes did not exceed $\pm 2.5\%$ with respect to the reference phoneme distribution for the Polish language, i.e. the phoneme balance comparable to this obtained for the German language [5].

Figure 2 depicts the data for the list set optimised for the binary scoring method[1] . The intelligibility functions for word-based scoring method are generally less steep than those obtained for binary scoring. The average SRT values for the lists lie with the range from -7.6 dB to -7.2 dB (mean $\mathrm{SRT} = -7.5$ dB). The average steepness falls in the range from 24.9%/dB to 28.9%/dB (mean steepness $S_{50} = 26.7$%/dB). The composed sentence materials are characterised by very similar phonemic content and the scattering of intelligibility data at SRT point is very low. It should be stressed that the psychometric function parameters characterizing the developed Polish sentence materials are different than those for other languages. For example, the German test, for word-based scoring, produced $\mathrm{SRT} = -6.23$ dB and $S_{50} = 19.20$%/dB [5] and the Dutch test, for binary-scoring, produced $\mathrm{SRT} = -4.07$ dB and $S_{50} = 16.30$%/dB [14]. These inconsistencies might be due to some linguistic, structural and acoustical differences be-

[1] An application enabling assessment of sentence intelligibility in noise for an exemplary sentence list is available at www.amu.edu.pl/∼hearcom. For a sake of convenience, an adaptive procedure for SRT estimation was employed.

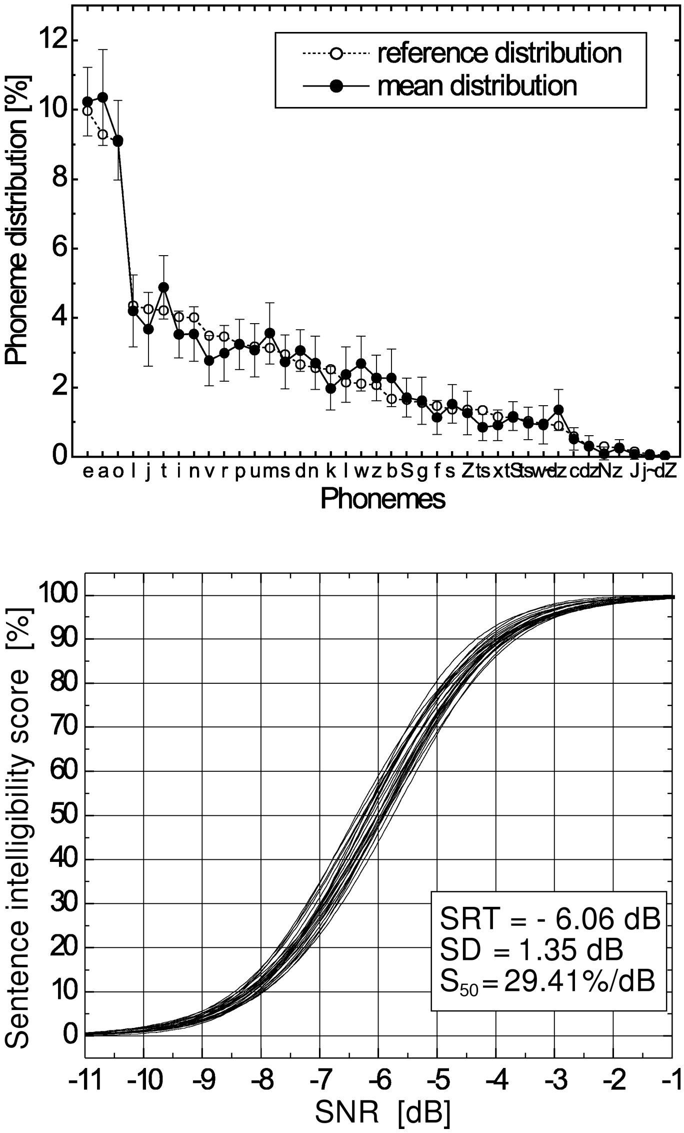

Fig. 2. The data set for the binary scoring. The top panel presents mean phoneme distribution (filled circles) averaged across 25 sentence lists and corresponding standard deviations (vertical bars) and reference distribution for the Polish language (open circles). The bottom panel depicts a juxtaposition of mean psychometric functions for 25 sentence lists.

tween these languages. Moreover, differences in magnitude of temporal fluctuations of the maskers might have influenced parameters of the psychometric functions obtained in different laboratories. The babble speech noise, used in the present study, is characterised by a relatively large envelope fluctuations. Therefore, it could have interrupted an audibility of single phonemes of utterances presented at low SNRs, that brought about reduction in speech intelligibility. Furthermore, a preliminary measurement suggested that the masker co-modulation effect might influence the psychometric function parameters.

Summarizing one can say that the main purpose of this study aimed at developing of reliable Polish sentence materials for accurate intelligibility measurements under noisy conditions, has been achieved. In order to accurately determine an individual SRT value in a real clinical situation, it is sufficient to collect the intelligibility scores for several randomly selected lists presented at different SNR ratios. If in a real clinical practice "typed" responses were not recorded (like in the traditional speech CVC audiometry), time of the data collection for one patient would decrease to about 10 minutes.

Acknowledgment

This paper is supported by grants from the European Union FP6, Project 004171 HEARCOM and the State Ministry of Science and Higher Education.

References

[1] Brachmański S., Staroniewicz P., *Fonetyczna struktura materiału testowego stosowanego w subiektywnych pomiarach jakości mowy*, [in:] Speech and Language Technology, Poznań 1999.

[2] Grocholewski S., *Statystyczne podstawy systemu ARM dla języka polskiego*, Wydawnictwo Politechniki Poznańskiej, **362** (2001).

[3] Hagerman B., *Sentences for testing speech intelligibility in noise*, Scand. Audiol., **11**, 79–87 (1982).

[4] Kalikow D., Stevens K., Elliot L., *Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability*, J. Acoust. Soc. Am., **61**, 1337–1351 (1977).

[5] Kollmeier B., Wesselkamp M., *Development and evaluation of a sentence test for objective and subjective speech intelligibility assessment*, J. Acoust. Soc. Am., **102**, 1085–1099 (1997).

[6] Nilsson M., Soli S., Sullivan J., *Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise*, J. Acoust. Soc. Am., **95**, 1085–1099 (1994).

[7] Peissig J., Kollmeier B., *Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners*, J. Acoust. Soc. Am., **101**, 1660–1670 (1997).

[8] Plomp R., Mimpen A. M., *Improving the reliability of testing the speech reception threshold for sentences*, Audiology, **18**, 43–53 (1979).

[9] Pruszewicz A., Demenko G., Richter L., Wika T., *New articulation lists for speech audiometry*. Part II, Otolaryngol. Pol., **48**, 56–62 (1994).

[10] Pruszewicz A., Demenko G., Richter L., Wika T., *New articulation lists for speech audiometry*. Part I, Otolaryngol. Pol., **48**, 50–55 (1994).

[11] Smoorenburg G. F., *Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram*, J. Acoust. Soc. Am., **91**, 421–437 (1992).

[12] Steffen–Batogowa M., *Automatyzacja transkrypcji fonematycznej tekstów polskich*, PWN, Warszawa 1975.

[13] Studebaker G., Sherbecoe R., McDaniel D. M., Gwaltney C. A., *Monosyllabic word recognition at higher-than-normal speech and noise levels*, J. Acoust. Soc. Am., **105**, 2431–2444 (1999).

[14] Versfeld N. J., Daalder L., Festen J. M., Houtgast T., *Method for the selection of sentence material for efficient measurement of the speech reception threshold*, J. Acoust. Soc. Am., **107**, 1671–1684 (2000).

[15] Wagener K., Josvassen J.L, Ardenkjaer R., *Design, optimization, and evaluation of a Danish sentence test in noise*, Journal of International Audiology, **42**, 10–17 (2005).

[16] Wouters J., Literie L., van Wieringen A., *Speech intelligibility in noisy environments with one- and two-microphone hearing aids*, Audiology, **38**, 91–98 (1999).